CONTRIBUTIONS TO VARIABLE SELECTION FOR MEAN MODELING AND VARIANCE MODELING IN COMPUTER EXPERIMENTS

A Thesis
Presented to
The Academic Faculty

by

Nagesh Adiga

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2012

COPYRIGHT 2012 BY NAGESH ADIGA

CONTRIBUTIONS TO VARIABLE SELECTION FOR MEAN MODELING AND VARIANCE MODELING IN COMPUTER EXPERIMENTS

Approved by:

Dr. Jye-chie Lu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jianjun Shi
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. C. F. Jeff Wu, Co-advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Martha Grover
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: Jan 11, 2012

To my parents,

Sooryanarayana Adiga and Girija Adiga,

and my beloved wife

Deepthi,

for their support, inspiration and encouragement

during my challenging journey

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all who have influenced, supported, and inspired my work in many ways.

First, I would like to express my sincere gratitude to my advisor, Professor. Jye-Chyi Lu for his guidance, encouragement, and hearty support during my doctoral program. He cared about both my personal and academic well being. He has not only been my academic advisor, but also a great mentor for my graduate student life.

I am extremely thankful to my co-advisor, Professor C. F. Jeff Wu for his guidance and support in my research, and wholehearted support during my studies. He has supporting me all through my doctoral program.

I would like to thank Dr. Tirthankar Dasgupta, and Dr. Roshan Joseph Vengazhiyil who have been extensively supporting and guiding me in my research and academics, and also been my mentor. They have constantly encouraging me in achieving my goal.

I am very thankful to my academic seniors Dr. Ying Hung, Dr. Xinwei Deng, Dr. Lulu Kang who shared space, time and knowledge with me at Georgia Tech. I thank my academic colleagues Dr. Sungil Kim, Hinkyeol Woo, Huizhi Xie, Shan Ba, Chia-Jung Chang and Dr. Ran Jin who have helped me during my doctoral program. I would like to thank all staff members of ISyE, especially Pamela Morrison and Mark Reese foe their kind support and help all stages of my graduate life at Ge,orgia Tech.

I would like to thank my wonderful friends in Atlanta. My roommates over the years, Manoj Agrawal, Dr. Shreekrishna, Dr. Dhaval Bhandari and Ajay Madhavan who have been cordial to me every day. I also thank my close friends Dr. Sandeep Kakumanu,

Dr. Kiruthika Devaraj, Yogish Gopala, Dr. Saikumar Thummurulu, Balachandra Suri, Ranjini Vaidyanathan, Yash Kochar, Saptarshi Ramanathan, Dr. Nischinth Rajmohan Manali Tare, Sravani, Dipti for all the great times spent together.

I would like to thank all my teachers at Indian Statistical Institute Kolkata, especially Dr. Biswabrata Pradhan who inspired me to pursue my doctoral program. I thank my mentors during my undergraduate studies in Karnataka Regional Engineering College, Surathkal, Dr. Suresh Hebbar and Dr. H.S.Y. Shasthry who have inspired me to pursue my passion in Statistics.

Last, but not the least, my heartfelt appreciation and gratitude goes to my family, especially my parents Soorya Narayana Adiga and Girija Adiga, my wife Deepthi, my sister Vani Herle, and my in-laws who have been supporting, encouraging and patient all through my doctoral program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis consists of two parts. The first part reviews a Variable Search, a variable selection procedure for mean modeling. The second part deals with variance modeling for robust parameter design in computer experiments.

In the first chapter of my thesis, Variable Search (VS) technique developed by Shainin (1988) is reviewed. VS has received quite a bit of attention from experimenters in industry. It uses the experimenters' knowledge about the process, in terms of good and bad settings and their importance. In this technique, a few experiments are conducted first at the best and worst settings of the variables to ascertain that they are indeed different from each other. Experiments are then conducted sequentially in two stages, namely swapping and capping, to determine the significance of variables, one at a time. Finally after all the significant variables have been identified, the model is fit and the best settings are determined.

The VS technique has not been analyzed thoroughly. In this report, we analyze each stage of the method mathematically. Each stage is formulated as a hypothesis test, and its performance expressed in terms of the model parameters. The performance of the VS technique is expressed as a function of the performances in each stage. Based on this, it is possible to compare its performance with the traditional techniques.

The second and third chapters of my thesis deal with variance modeling for robust parameter design in computer experiments. Computer experiments based on engineering models might be used to explore process behavior if physical experiments (e.g. fabrication of nanoparticles) are costly or time consuming. Robust parameter design

(RPD) is a key technique to improve process repeatability. Absence of replicates in computer experiments (e.g. Space Filling Design (SFD)) is a challenge in locating RPD solution. Recently, there have been studies (e.g. Bates et al. (2005), Chen et al. (2006), Dellino et al. (2010 and 2011), Giovagnoli and Romano (2008)) of RPD issues on computer experiments. Transmitted variance model (TVM) proposed by Shoemaker and Tsui. (1993) for physical experiments can be applied in computer simulations. The approaches stated above rely heavily on the estimated mean model because they obtain expressions for variance directly from mean models or by using them for generating replicates. Variance modeling based on some kind of replicates relies on the estimated mean model to a lesser extent. To the best of our knowledge, there is no rigorous research on variance modeling needed for RPD in computer experiments.

We develop procedures for identifying variance models. First, we explore procedures to decide groups of pseudo replicates for variance modeling. A formal variance change-point procedure is developed to rigorously determine the replicate groups. Next, variance model is identified and estimated through a three-step variable selection procedure. Properties of the proposed method are investigated under various conditions through analytical and empirical studies. In particular, impact of correlated response on the performance is discussed.

CHAPTER I

ANOTHER LOOK AT DORIAN SHAININ'S VARIABLE SEARCH TECHNIQUE

1.1 Introduction

Dorian Shainin, a well known quality consultant, developed a quality improvement program popularly known as the *Shainin system*. Shainin system has been reported to be useful to many industries. Among several new tools and techniques proposed by Shainin (Steiner, MacKay and Ramberg 2008), Variable Search (VS) is one which has received quite a bit of attention from researchers and practitioners. Described as "The Rolls Royce of Variance Reduction" by Bhote and Bhote (2001), VS technique (Shainin 1986, Shainin and Shainin 1988) can be described as a sequential screening process used to identify the key factors and their settings to optimize the response of a system by making use of available engineering knowledge. Verma et al. (2004) used some real life numerical examples to demonstrate the superiority of VS over more traditional methods.

It is worth mentioning that another Shainin method called component search (CS) has been found to be quite popular among engineers. CS (Shainin and Shainin 1988) is typically used when units can be disassembled and reassembled without damage or change to any of the components or subassemblies, with the objective of comparing families of variation defined by the assembly operation and individual components. CS has a stage called component swapping from which VS was derived. See Steiner et al. (2008) for a detailed description of CS.

Ledolter and Swersey (1997) critically examined the VS method and argued via an example involving seven factors, that a fractional factorial design is generally a

better alternative compared to the VS method. In this article, we carry out a more in-depth analysis of Shainin's VS procedure by (a) summarizing the key properties of the VS design and investigating its suitability for factor screening, (b) identifying the statistical inference procedures associated with each step of the VS design, (c) combining the test procedures at individual steps to obtain a general expression for the probability of correct identification of active factors, and (d) examining the sensitivity of the VS method against correctness of engineering assumptions and varying levels of noise.

The remainder of this chapter of the thesis is organized as follows. In Section 1.2, we provide a stage-by-stage description of the VS methodology. In Section 1.3, we discuss the salient properties of the VS design. Section 1.4 deals with the statistical inference procedures associated with individual stages of VS, and their impact on the overall process. In Section 1.5, we study the sensitivity of the VS method against correctness of engineering assumptions and varying levels of noise. Some concluding remarks are presented in Section 1.6.

1.2 An Overview of Shainin's Variable Search

Here we explain the VS method with a hypothetical example with seven factors (and data) presented in Table 1.1. The method consists of four stages, and assumes that the following are known: (i) the order of importance of the factors under investigation, and (ii) the best and worst settings for each of the factors.

In **Stage 1**, the suspect factors are ranked in descending order of perceived importance. Two levels are assigned to each factor: "best" (+) level and "worst" (−) level. Assume that the objective is to maximize the response, i.e., larger values of response are preferred. The algorithm starts with two experimental runs: one with all factors at their best levels and the other with all factors at their worst levels. These two runs are then replicated thrice in random order (runs 1 to 6 in Table 1.1). These

responses are used to test if there is a statistically significant difference between these two settings. Median and range for the three replications for the two experiments are computed. Denote these medians by M_b and M_w, and the ranges by R_b and R_w respectively, where the suffixes b and w indicate best and worst settings respectively. We then compute $R_m = (M_b - M_w)/R_{avg}$, where $R_{avg} = (R_b + R_w)/2$ denotes the average range. Note that R_{avg}/d_2 is an unbiased estimator of the underlying normal error standard deviation σ, where $d_2 = 1.693$ for a sample size of 3. A value of R_m greater than 1.07 (or, alternatively 1.25 as recommended by Bhote (2001)) suggests the presence of at least one active factor and prompts the experimenter to move on to the next stage of the algorithm. In the example given in Table 1.1, R_{avg}= (5+7)/2 = 6 and $M_b - M_w$ = 451-66 = 385, so that $R_m = 64.17$.

In **Stage 2**, Shainin specifies confidence intervals for the mean response corresponding to the "best" and "worst" situations based on t-distribution with four degrees of freedom as $M_b \pm 2.776 R_{avg}/d_2$ and $M_w \pm 2.776 R_{avg}/d_2$ respectively, where $d_2 = 1.693$ since the sample size is three. These confidence intervals are used in the later stages to determine significance of factors or groups of factors. In our example, the confidence intervals corresponding to the "best" and "worst" settings are computed as (441.2, 460.8) and (56.2, 75.8) respectively.

Stage 3, also called *Swapping*, is used to identify active factors by switching levels of each factor one at a time. Ideally, swapping should start from the most important factor and end with the least important factor. Assume, without loss of generality, that factor 1 is most important. Swapping of factor 1 is performed in the following way. The first run is conducted with factor 1 at "best" level and all the other factors at their "worst" levels. One more run is conducted with all these factor levels reversed. Factor 1 is declared *inert* (or insignificant) if both the response values are within the confidence intervals derived in stage 2, and *active* (or significant) otherwise. Similarly, swapping is performed with each other factor in order of perceived importance until

Table 1.1: Example of variable search design

Run	1	2	3	4	5	6	7	y	Confidence interval	Remark
1	+	+	+	+	+	+	+	448		
2	+	+	+	+	+	+	+	453		
3	+	+	+	+	+	+	+	451		
4	-	-	-	-	-	-	-	63		
5	-	-	-	-	-	-	-	66		
6	-	-	-	-	-	-	-	70		
	Swapping for Factor 1									
7	-	+	+	+	+	+	+	350	(441.2, 460.8)	1 is active
8	+	-	-	-	-	-	-	104	(56.2, 75.8)	1 is active
	Swapping for Factor 2									
9	+	-	+	+	+	+	+	324	(441.2, 460.8)	2 is active
10	-	+	-	-	-	-	-	249	(56.2, 75.8)	2 is active
	Capping runs (Factors 1 and 2)									
11	+	+	-	-	-	-	-	392	(441.2, 460.8)	Capping run unsuccessful
12	-	-	+	+	+	+	+	106	(56.2, 75.8)	Capping run unsuccessful
	Swapping for Factor 3									
13	+	+	-	+	+	+	+	403	(441.2, 460.8)	3 is active
14	-	-	+	-	-	-	-	96	(56.2, 75.8)	3 is active
	Capping runs (Factors 1, 2 and 3)									
15	+	+	+	-	-	-	-	443	(441.2, 460.8)	Capping run successful
16	-	-	-	+	+	+	+	60	(56.2, 75.8)	Capping run successful

two active factors are found. Once two active factors have been identified, we move to the next stage.

Stage 4, called *Capping*, is used to check whether there are still more active factors to be identified (apart from the already identified ones). Two runs are conducted to confirm this. In the first run, all the factors identified active are set at their "best" levels and all the other factors at their "worst" levels. In the second run, all these levels are reversed. If the two responses from these two trials lie within the confidence intervals computed at stage 2, it is concluded that all the active factors have been identified successfully. Otherwise, one needs to go back to the swapping stage and search for some more active factors.

Swapping and capping runs are successively conducted till a capping run is "successful", which means there are no more active factors to be identified. In the example in Table 1.1, swapping of factor 1 (runs 7-8) and factor 2 (runs 9-10) declare these two factors as active. The follow-up capping run with these two factors (runs 11-12) is unsuccessful, which leads to the conclusion that there are possibly more active factors. Swapping of factor 3 (runs 13-14) declares it as active. Finally capping of

factors (1,2,3) is successful, and leads to termination of the VS process.

1.3 Properties of Variable Search Design

Ledolter and Swersey (1997) discussed some properties of the VS design and compared it with a fractional factorial design using the second order model

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + (x_1x_2)\beta_{12} + (x_1x_3)\beta_{13} + (x_2x_3)\beta_{23} + \epsilon, \qquad (1.1)$$

where $\epsilon \sim N(0, \sigma^2)$. Assuming that out of seven factors $(1, 2, \ldots, 7)$ under investigation, three (1,2,3) are active through model (1.1), they argued that a 2^{7-3} fractional factorial design of resolution IV (Wu and Hamada 2000, Ch. 4) with defining relations $5 = 123$, $6 = 124$ and $7 = 234$ is superior to the VS design in terms of estimation efficiency. Note that the best possible VS design in this case with the correct conclusion is the one shown in Table 1.1.

However, it can be seen that, in this case, the VS design may have some advantages over the 2^{7-3} design in the context of early identification of active effects, provided the experimenter's knowledge regarding the relative importance of factors and their "best" and "worst" levels is perfect. Assume that the error variance σ^2 is sufficiently small to ensure that the statistical tests of hypotheses are powerful enough to guarantee the detection of significant effects both by the VS design and the fractional factorial design. Then, usual analysis of data obtained from the 16-run 2^{7-3} design will declare main effects $1, 2, 3$ and two factor interactions (2fi's) $12, 13$ and 23 as significant. However, the 2fi's 12, 13 and 23 are aliased with other 2 fi's, i.e., $12 = 35 = 46$, $13 = 25$, and $23 = 15 = 47$. Whereas the 2fi's 46 and 47 can be ruled out using the effect heredity principle (Wu and Hamada 2009, Ch 5), the 2 fi's 35, 25 and 15 cannot, as one of the parent factors of each is significant. Thus, one would need additional orthogonal runs (at least 4) to disentangle these effects and identify the *true* active factors.

This problem will not arise for the VS design because the swapping and capping

runs permit simultaneous testing of main effects and interactions associated with a single factor or a group of factors (this will be discussed in detail in the next section), thereby enabling the experimenter to completely rule out the presence of main effects or interactions of groups of effects. Therefore, in order to unambiguously identify the active factors, the fractional factorial may actually need more runs than the best possible VS design. This relative advantage of the VS design will be more predominant in this example if the model (1.1) contains a three-factor interaction (3fi) term 123. In this case, regular analysis of the fractional factorial will identify the main effects 1, 2, 3 and 5 as significant, which will create more confusion regarding construction of orthogonal follow-up runs. The VS design will, however, remain the same if the sequence of runs is correct.

Now, consider the case where the underlying true model is still the same as (1.1), but the number of factors under investigation is 20. Under the same assumption as before, the VS design will still be able to reach the correct conclusions in 16 runs, while a 16-run fractional factorial (2^{20-16}) clearly cannot be constructed. One should keep in mind, though, that the assumption of a very accurate level of process knowledge that leads to conducting swapping runs for the 3 active factors (out of 20) first is a very strong one.

Thus, whereas the VS design has some properties (e.g., estimation efficiency) that make it inferior to a comparable fractional factorial design, it also has certain properties that may give it an edge in *specific* situations. Apart from such plausible technical advantages, it should be noted that the VS method has certain *practical advantages*. The sequential nature of the experiments in the VS design permits the experimenter to obtain partial information at each stage of the experiment, whereas in fractional factorial or orthogonal array designs, one needs to wait for the experiments to be completed. The simplicity of the VS technique is appealing to the experimenters. However, the VS method has some *practical drawbacks* as well. It will not be efficient if

applied to a new process, or to a process with no prior information. Also the frequent change of settings will be expensive if the levels of factors are hard to change. By contrast, restricted randomization can be used in fractional factorial designs. Note that one critical disadvantage that may overshadow the practical benefits accrued from the sequential nature of experiments in the VS design (as stated earlier) is the presence of block effects in the form of uncontrollable factors drifting over time. However, since the experiment has variable block sizes, incorporating block effects into the analysis of VS designs appears to be quite non-trivial. Therefore, in the paper, we assume no block effect and leave the analysis with block effect as future research.

In the following subsections we study some properties of the VS design in terms of run size and estimation efficiency. The proofs of all the results stated in this section are given in the Appendix A.

1.3.1 Run Size of the VS Design

If we assume that each stage of the VS design would result in correct conclusion (the probability of which will be explored in Section 1.4), then the number of runs will depend on the ordering of the factors according to perceived importance. Clearly, the best possible scenario would be one where all the active factors are explored first, and the worst possible scenario would occur when the last factor to be explored is an active factor. The result in (1.2) is useful to compute the smallest possible, largest possible, and expected run length of a VS design. Suppose that the VS design identifies m active factors out of k factors under investigation, where $1 \leq m \leq k$. Then, the total number of runs N of the VS design satisfies

$$\begin{cases} N = 2(k+3), & \text{if} \quad m = 1, \\ 4(m+1) \leq N \leq 2(k+m)+4, & \text{if} \quad m > 1. \end{cases} \tag{1.2}$$

For example, if the VS design identifies 3 out of 7 factors under investigation as active (as in the example given in Table 1.1), the minimum and maximum number of runs of the VS design will be 16 and 24 respectively. The following result is helpful to compute the expected run size for the VS design under the assumption of a random ordering of factors chosen for swapping runs.

Theorem 1. *Suppose p $(2 \leq p \leq k)$ out of k factors being investigated are actually active. Assume that*

(i) Each swapping and capping run will lead to a correct conclusion.

(ii) There is a complete lack of engineering knowledge regarding relative importance of factors, which means any permutation of $(1, 2, \ldots, k)$ is equally likely to occur while investigating the k factors one-by-one.

Then, the total number of runs in the VS design is a discrete random variable N that has the following probability mass function: $Pr\{N = 4(p+1) + 2j\} = p\binom{k-p}{j}/(p+1)$ and its expectation is given by $E(N) = 4(p+1) + 2(k-p)p/(p+1)$.

Using Theorem 1, it is seen that in the complete absence of knowledge about the relative importance of factors, the expected number of runs of a VS design in the example discussed in Section 1.2 with $k = 7$ and $p = 3$ is 22. A perfect knowledge of the relative importance will reduce the number of required runs to 16. An investigation with $k = 20$ factors will need, on average, about 42 runs if $p = 3$. The significant savings of runs that can be achieved by using VS design over comparable fractional factorial design is therefore quite evident.The probability mass functions of N for different values of k and $p = 3$ are shown in Figure 1.1.

1.3.2 Estimation of main effects from VS Design

We begin this discussion by noting that the VS design is neither a design with a fixed number of runs nor an orthogonal array. Assume that p out of k factors under

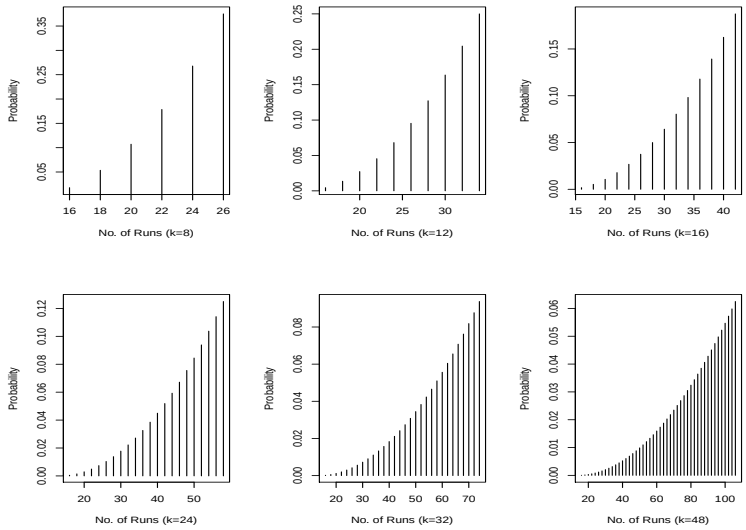


Figure 1.1: Probability mass function of N for $k = 8, 12, 16, 24, 32$ and 48, when $p = 3$.

investigation are active, and the following model describes their relationship with the response y:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i<j} \sum_{j=1}^{p} \beta_{ij} x_i x_j + \ldots + \beta_{12\ldots p} \;\; x_1 x_2 \ldots x_p + \varepsilon, \tag{1.3}$$

where $x_i = -1$ or $+1$ according as the worst or best level of factor i is used, and $\epsilon \sim N(0, \sigma^2)$. Further, assume that the p active factors are correctly identified by the VS procedure.

Shainin recommended estimation of main effects of active factors from the $(2p + 2) \times p$ submatrix of the VS design that consists of two stage I runs and p pairs of swapping runs for the active factors. We shall denote this matrix by $\mathbf{X}_p$ in all subsequent discussions. Table 1.2 shows matrix $\mathbf{X}_7$. Shainin suggested estimating the main effect of factor i for $i = 1, 2, \ldots, p$ (defined as twice the regression coefficient β_i in model (1.3)) by comparing its swapping runs to the two corresponding stage I runs. For example, the regression coefficients β_1 and β_2 are estimated as

$$\hat{\beta}_1 = (y_1 - y_2 - y_3 + y_4)/4, \tag{1.4}$$

$$\hat{\beta}_2 = (y_1 - y_2 - y_5 + y_6)/4. \tag{1.5}$$

Table 1.2: Model matrix for estimation from VS design

Run	1	2	3	4	5	6	7	y
1	+	+	+	+	+	+	+	y_1
2	-	-	-	-	-	-	-	y_2
3	-	+	+	+	+	+	+	y_3
4	+	-	-	-	-	-	-	y_4
5	+	-	+	+	+	+	+	y_5
6	-	+	-	-	-	-	-	y_6
7	+	+	-	+	+	+	+	y_7
8	-	-	+	-	-	-	-	y_8
9	+	+	+	-	+	+	+	y_9
10	-	-	-	+	-	-	-	y_{10}
11	+	+	+	+	-	+	+	y_{11}
12	-	-	-	-	+	-	-	y_{12}
13	+	+	+	+	+	-	+	y_{13}
14	-	-	-	-	-	+	-	y_{14}
15	+	+	+	+	+	+	-	y_{15}
16	-	-	-	-	-	-	+	y_{16}

It is easy to see that the above estimators are unbiased, have a pairwise correlation of 0.5, and have the same standard error of 0.5σ. Ledolter and Swersey (1997) are particularly critical about this estimation procedure, owing to its large standard error (the standard error of $\hat{\beta}_i$ estimated from a 16-run factorial design would be 0.25σ) and correctly argue that a least squares estimator would be a better choice as it is statistically more efficient. For example, for $p = 7$, we find that the standard error of the least squares estimator $\hat{\beta}_i$ is 0.33σ, and each pair of estimated effects $\hat{\beta}_i, \hat{\beta}_j$ have a correlation of approximately -0.14. We shall now see that the least squares estimators obtained this way have some interesting properties that are summarized in Lemma 1 and Theorem 2.

Lemma 1. Let $\mathbf{x}_i$, $i = 1, 2, \ldots, p$, denote the ith column of $\mathbf{X}_p$. Then $\mathbf{x}_i$ is orthogonal to the interaction column generated by $\mathbf{x}_i$ and any other column $\mathbf{x}_j$, $j = 1, \ldots, p$, $j \neq i$.

Theorem 2. *Assume that* $\mathbf{y}$*, the* $(2p + 2) \times 1$ *vector of responses, depends on the* p *active factors through model (1.3), and let* $\boldsymbol{\beta}_{main} = (\beta_1, \ldots, \beta_p)'$.

(i) Then, the least squares estimator of $\boldsymbol{\beta}_{main}$ *given by*

$$\hat{\boldsymbol{\beta}}_{main} = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{y} \tag{1.6}$$

satisfies the following:

$$E(\hat{\boldsymbol{\beta}}_{main}) = \boldsymbol{\beta}_{main} \text{ if all three and higher order interactions in model (1.3) are absent}$$

$$(ii)\sigma_p^2 = \frac{p^2-4p+7}{8(p^2-3p+4)}\sigma^2 \text{ for } i=1,\ldots,p, \tag{1.7}$$

$$\rho_p = -\frac{p-3}{p^2-4p+7} \text{ for } i,j=1,\ldots,p, i\neq j, \tag{1.8}$$

where σ_p^2 and ρ_p denote the variance and pairwise correlation of estimated main effects when there are p active factors.

From (1.7) and (1.8), we have that $\sigma_p^2 = 0.125\sigma^2$ for $p = 3$ and $\sigma_p^2 \to 0.125\sigma^2$ as $p \to \infty$. Also, $\rho_p = 0$ for $p = 3$ and $\rho_p \to 0$ as $p \to \infty$. Figure 1.3 shows plots of σ_p^2 (dotted curve in the left panel) and ρ_p (right panel) for $\sigma = 1$. From the above observations and Figure 1.3 , the results in (1.9)-(1.10) can easily be established. The variance σ_p^2 and the correlation coefficient ρ_p satisfy the following inequalities

$$0.1071\sigma^2 \leq \sigma_p^2 \leq 0.1250\sigma^2, \tag{1.9}$$

$$-0.167 \leq \rho_p \leq 0. \tag{1.10}$$

The lower and upper bounds in (1.9) and (1.10) are attained for $p = 5$ and $p = 3$ respectively.

The findings from Lemma 1 and Theorem 2 can be summarized as follows:

1. Least squares estimators of main effects of active factors obtained from stage-I and swapping runs enumerate are *unbiased* and uncorrelated with estimators of 2fi's under the assumption of negligible three factor interactions.

2. The standard error of these estimators remains almost invariant (varying from $\sqrt{.1071}\sigma = 0.33\sigma$ to $\sqrt{.125}\sigma = 0.35\sigma$) with respect to the number of active factors.

3. The estimators are uncorrelated only if $p = 3$. For $p \geq 3$, they have a small negative correlation, which has the largest magnitude (-0.167) for $p = 5$.

Comparison of efficiency with folded-over Plackett-Burman type screening designs

A large class of two-level orthogonal array designs with run size $N = 4n$ (where n is a positive integer) given by Plackett and Burman (1946) have been used as screening designs. Box and Hunter (1961) demonstrated that resolution IV designs can be obtained by folding over such designs. Miller and Sitter (2001) further studied these designs for $N = 12$ and proposed an analysis strategy.

Note that when the number of factors p satisfies $p+1 = 4n$, a folded-over version of the Plackett-Burman (FPB) design is of order $(2p+2)\times p$, and is therefore comparable to the VS design matrix $\mathbf{X}_p$. The variance of $\hat{\beta}_i$ estimated from the FPB design with p factors is $\widetilde{\sigma}_p^2 = \sigma^2/(2p+2)$. Comparing this with (1.7), the relative efficiency of the FPB design with respect to that of the VS design can be obtained as

$$e_p = \frac{\widetilde{\sigma}_p^2}{\sigma_p^2} = \frac{1}{4}\big(p - \frac{p-7}{p^2-3p+4}\big). \tag{1.11}$$

From (1.11), we find that $e_p = 1$ when $p = 3$, i.e., both the designs are equally efficient with respect to estimation of main effects. However, for $p = 7$, we have $e_p = 7/4$, which means that the VS design is almost half as efficient as the FPB design. Further, $e_p \to \infty$ as $p \to \infty$, as seen from Figure 1.2. A comparison of variance of $\hat{\beta}_i$ obtained from the FPB and VS designs is shown in the left panel of Figure 1.3 for different values of p.

Although the above discussion clearly establishes the superiority of FPB designs over VS designs with respect to estimation efficiency, two important points should be kept in mind. First, the FPB design exists only when $p = 4k - 1$. For example, when $p = 4, 5$ or 6, an orthogonal design comparable to the VS design does not exist. Second, and more importantly, the FPB design would usually include all of the $k(> p)$ factors under investigation, whereas the VS design matrix $\mathbf{X}_p$ corresponds only to the p factors that are screened out as active. Thus the above efficiency comparison may not always be meaningful. Further, when the number of factors under investigation is large, the VS method will have a clear-cut advantage in terms of run size as described

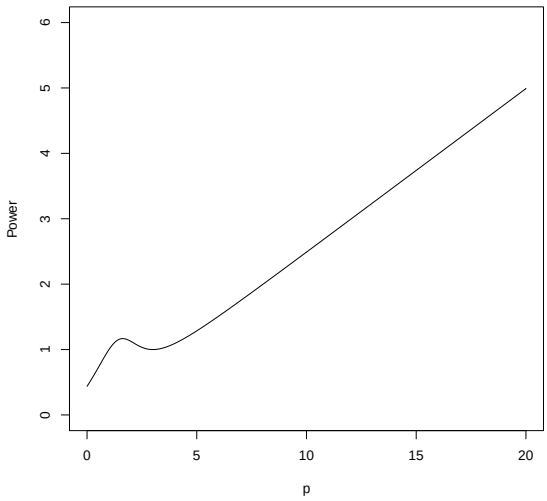


Figure 1.2: Plot of e_p

earlier in Section 1.3.1.

1.3.3 Estimation of interaction effects

So far we have discussed only the estimation of main effects. In the VS design, the 2fi's are aliased with one another, and as observed by Ledolter and Swersey (1997), the swapping runs for factor i permit estimation of the sum of the 2fi's associated with that factor (more discussion on this in Section 4.1.2). We now discuss some properties of the VS design, summarized in the following two Theorems (3 and 4), that helps us to devise a strategy to obtain unconfounded estimates of 2fi's (and higher order interaction effects, if they exist) when the number of active factors do not exceed 4. It is well known that good screening designs should have projection properties, for which Box and Tyssedal (1996) gave the following definition.

Definition. An $N \times k$ design $\mathcal{D}$ with N runs and k factors each at 2 levels is said to be of projectivity P if every subset of P factors out of the possible k contains a complete 2^P full factorial design, possibly with some points replicated. The resulting design will then be called a (N, k, P) screen.

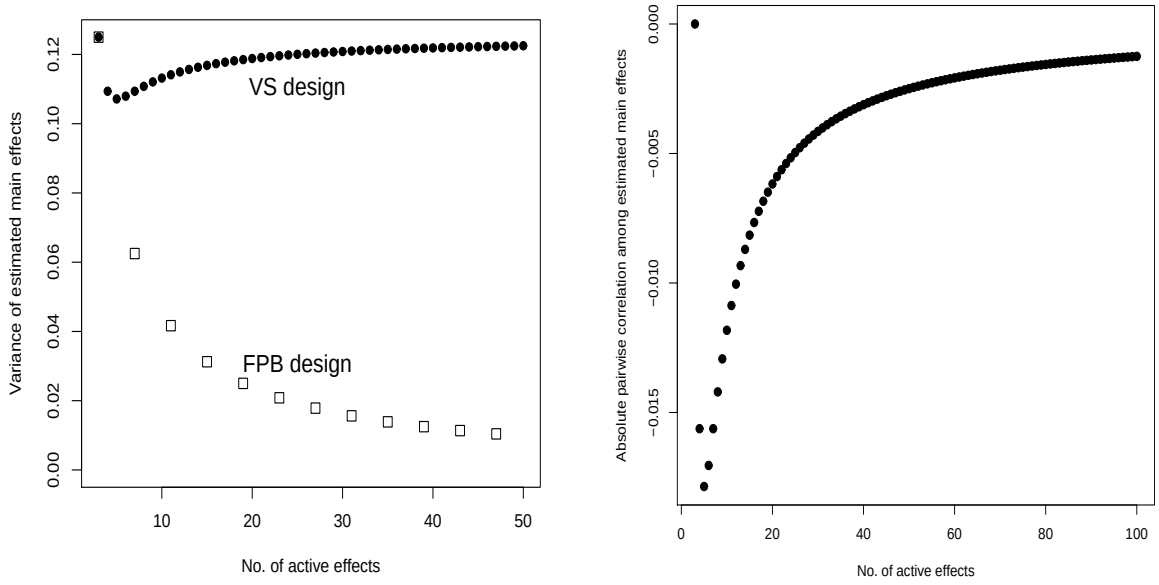


Figure 1.3: Plots of σ_p^2 (left panel) and ρ_p (right panel) for $\sigma = 1$.

The above definition refers to designs that are orthogonal arrays. However, extending the definition of (N, k, P) screen to all $N \times k$ designs that are not necessarily orthogonal arrays, the following Results can easily be established for a VS design.

Theorem 3. *Consider an $N \times m$ submatrix $\mathbf{D}$ of a VS design matrix that consists of the columns corresponding to m (≥ 3) factors identified as active. Then,*

(i) submatrix $\mathbf{D}$ is a $(N, k, 3)$ screen, i.e., has projectivity 3.

Further, when $m = 4$ with four active factors A, B, C and D identified by VS,

(ii) submatrix $\mathbf{D}$ contains a 2_{IV}^{4-1} fractional factorial design in these four factors with the defining relation $I = -ABCD$.

(iii) The above fractional factorial design is the largest *orthogonal array that is contained in the submatrix $\mathbf{D}$.*

(iv) It is possible to construct a 2^4 design in A, B, C and D by addition of just four more runs $(+--+)$,$(+---+)$,$(-++-)$ and $(-+-+)$ to the VS design.

From Theorem 3, it follows that for $p = 3$, the VS design permits unbiased and independent estimation of *all* factorial effects from the 2^3 design that it contains. Note that the design matrix is precisely the one discussed earlier in the context of least squares estimation of main effects, that consists of two stage-I and the swapping runs. Any regression coefficient in model (1.3) estimated in this way from the VS design will have a standard error of $\sigma/\sqrt{8} = 0.35\sigma$. When $p = 4$, the VS design still permits independent estimation of the four main effects and three aliased sets of 2fi's from the 8-run fractional factorial design identified in Theorem 4(i) with a standard error of 0.35σ. The standard error of each main effect estimate can be slightly reduced to 0.33σ by using least squares estimation with the upper left 10×4 submatrix of the design matrix in Table 1.2; however, this will result in correlated estimates. A better strategy may be to conduct four additional runs as suggested in Theorem 4(iii) and estimate each factorial effect of every possible order with a standard error of 0.125σ.

1.4 Statistical Inference at Different Stages of Variable Search

The objective of this section is to understand the mechanism of hypothesis testing associated with the VS design, compute the probabilities of obtaining correct conclusions at different stages of the VS design (equations (14), (17) and (22)), and eventually utilize these results to compute the probability of correct screening of active factors if the order of investigation of factors is fixed. Readers who wish to skip the technical details may skip the derivations in the first subsection and move on to the next subsection, which is of more practical importance.

Assume that VS is being performed to identify the active factors from a pool of k potential factors $x_1, \ldots, x_k$. The best and worst levels of each factor are known and

represented by $+1$ and -1 respectively. The objective is to maximize the response y, which is related to the experimental factors through the following second-order model:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \varepsilon, \tag{1.12}$$

where $x_i = -1$ or $+1$ according as the worst or best level of factor i is used, and $\epsilon \sim N(0, \sigma^2)$. Note that, correct knowledge of the best and worst levels of each factor implies that $\beta_i > 0$ for $i = 1, \ldots, k$ in model (1.12). Next, we introduce the following notation:

- y_i^+ (y_i^-): Observed response when factor x_i is at '+' ('$-$') level and the remaining $k-1$ factors are at '$-$' ('+') level.
- Let $\mathcal{F} \subset \{1, 2, \ldots, k\}$ be a set with cardinality q. Define $y_{\mathcal{F}}^+$ ($y_{\mathcal{F}}^-$) as the observed response when all the q factors $x_i, i \in \mathcal{F}$ are at '+' ('$-$') level and the remaining $k-q$ factors are at '$-$' ('+') level.
- y^+ (y^-): Observed response when all the k factors are at '+' ('$-$') level.

In the following subsections, we describe the statistical tests of hypothesis associated with different stages of the VS design. To keep our computations tractable, we shall (i) use the large sample approximation for sample median, although the sample size in the VS design is only 3, and (ii) use sample standard deviation instead of sample range to estimate σ.

1.4.1 Power of Statistical Hypotheses Tested at Different Stages of VS and Probability of Correct Screening

1.4.1.1 Stage I

As described in Section 1.2, the median M_b of three realizations of y^+ are compared with the median M_w of three realizations of y^-.

It is easy to verify (see Appendix A for a detailed argument) the following:

- As observed by Ledolter and Swersey (1997), stage-I of VS with respect to model (1.12) is equivalent to testing the hypothesis $H_0 : \sum_{i=1}^{k} \beta_i = 0$ against $H_1 : \beta_i \neq 0$ for at least one $i = 1, 2, \ldots, k$. An appropriate rejection rule will be the following: reject H_0 at level α if

$$\frac{|M_b - M_w|}{\hat{\sigma}\sqrt{\pi/3}} > t_{4,\alpha/2}. \tag{1.13}$$

- The power (i.e., ability to detect presence of at least one active factor) of the test is given by

$$P_I = 1 - F_\delta(t_{4,\alpha/2}) + F_\delta(-t_{4,\alpha/2}), \tag{1.14}$$

  where $F_\delta(\cdot)$ denotes the cumulative distribution function of a non-central t distribution with 4 degrees of freedom and non-centrality parameter $\delta = (2\sum_{i=1}^{k} \beta_i)/(\sigma\sqrt{\pi/3})$.

Clearly, P_I is a monotonically increasing function of $(\sum_{i=1}^{k} \beta_i)/\sigma$ through the parameter δ. Figure 1.5 shows a plot of the power function when α is equal to 0.05. It is seen that the power is approximately 95% if $(\sum_{i=1}^{k} \beta_i)/\sigma \approx 2.52$ and 99% if $(\sum_{i=1}^{k} \beta_i)/\sigma \approx 3.07$. In other words, the power of the stage-I test is 95% and 99% if the sum of main effects $\sum_{i=1}^{k} \beta_i$ is respectively equal to 2.5 times and 3 times the error standard deviation σ. The plot of the power function, for α equal to 0.01 and 0.10 are shown in Figure 1.4 and Figure 1.6 respectively. It can be seen that the power of the test increases faster when α is higher and viceversa.

1.4.1.2 Swapping

The objective of the swapping stage it to identify active factors. Here, with reference to the experiments in stage I, the level of one particular factor is switched, and the difference in the response is observed. If this change is significant, then that factor is declared as active. Details of the hypothesis test associated with the swapping of

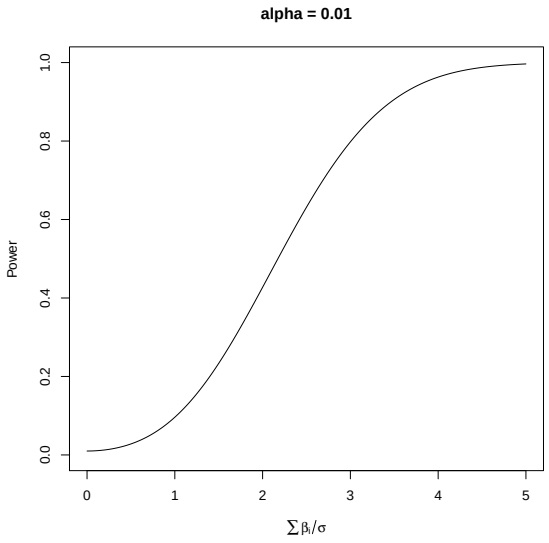


Figure 1.4: Power of Stage I test as a function of $\sum \beta_i/\sigma, \alpha = 0.01$.

factor x_i is described in the Appendix. The power of the test can be obtained as

$$P^i_{\text{swap}} = 1 - \{F_{\delta^+}(t_{4,\alpha/2}) - F_{\delta^+}(-t_{4,\alpha/2})\}\{F_{\delta^-}(t_{4,\alpha/2}) - F_{\delta^-}(-t_{4,\alpha/2})\}, \qquad (1.15)$$

where $F_{\delta^+}(\cdot)$ and $F_{\delta^-}(\cdot)$ denote the cumulative distribution function of a non-central t distribution with 4 degrees of freedom and non-centrality parameters $\delta^+ = (2\beta_i + 2\sum_{j\neq i}\beta_{ij})/(1.23\sigma)$ and $\delta^- = (2\beta_i - 2\sum_{j\neq i}\beta_{ij})/(1.23\sigma)$ respectively.

Thus, the power of the swapping phase is a function of β_i/σ and $\sum_{j\neq i}^{k}\beta_{ij}/\sigma$. Figure 1.7, Figure 1.8, and Figure 1.9 show contour plots of the power against β_i/σ and $\sum_{j\neq i}^{k}\beta_{ij}/\sigma$ when α is equal to 0.01, 0.05 and 0.10 respectively. The darker regions represent low powers. As expected, the power of the test is small when both β_i/σ and $|\sum_{j\neq i}^{k}\beta_{ij}|/\sigma$ are small. Also, as in Stage I, the power of the test increases faster when α is high, and viceversa.

Remarks:

1. The type-I error of the test described above will be $1-(1-\alpha)^2$. To ensure that the test has a pre-specified type-I error α, the levels of significance associated with the events A^+ and A^- should be adjusted.

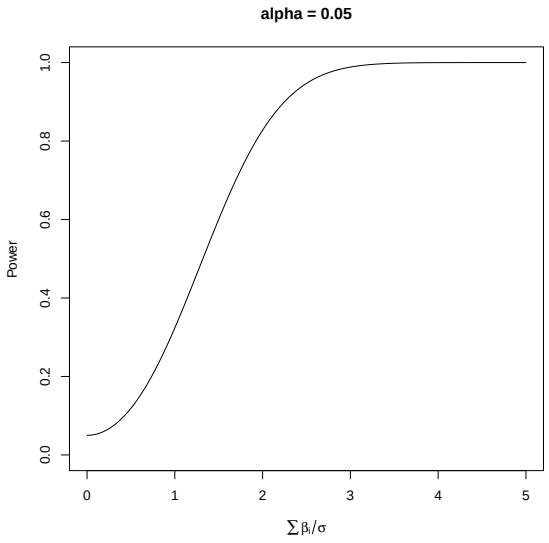


Figure 1.5: Power of Stage I test as a function of $\sum \beta_i/\sigma, \alpha = 0.05$.

2. Note that each pair of swapping runs for an active factor x_i leads to estimation of the main effect β_i and sum of all interactions $\sum_{j \neq i} \beta_{ij}$ that involve x_i. Define $u_i = (s^+(x_i) + s^-(x_i))/4$ and $v_i = (s^+(x_i) - s^-(x_i))/4$. Then, unbiased estimators of β_i and $\sum_{j \neq i} \beta_{ij}$ are given by

$$\hat{\beta}_i = u_i, \tag{1.16}$$

$$\sum_{j \neq i} \hat{\beta}_{ij} = v_i, \tag{1.17}$$

for $i = 1, \ldots, p$. The variance of $\hat{\beta}_i$ is given by $\frac{1}{8}(1 + \pi/6)\sigma^2$ or $0.19\sigma^2$. Note that this estimation procedure is the same as the one recommended by Shainin and discussed earlier in Section 3.2, except for the fact that six stage-I runs are used instead of two. This reduces the variance of the estimates, but inflates the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$.

3. Thus, in a nutshell, swapping allows combined testing of all effects involving a single factor, and hence is a useful tool to detect whether a single factor is "active" by conducting only two additional runs.

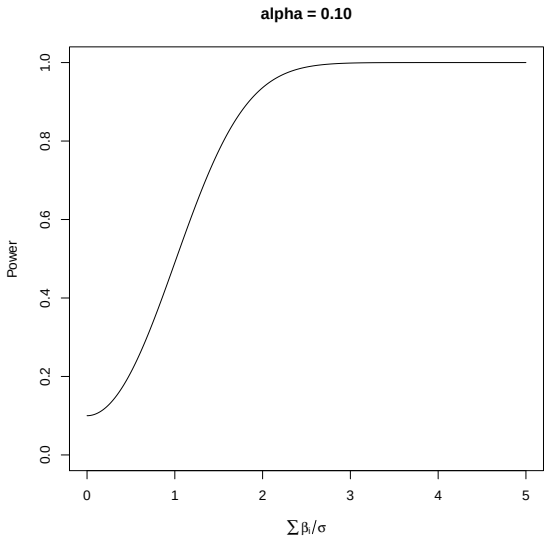


Figure 1.6: Power of Stage I test as a function of $\sum \beta_i/\sigma, \alpha = 0.10$.

1.4.1.3 Capping

The objective of the capping stage is to confirm whether all the active factors have been identified. Assume that q out of the k factors have been declared active in the swapping stage. Let $\mathcal{F}$ represent the set of indices of factors which have been declared active. Details of the hypothesis test associated with the capping of the factors in $\mathcal{F}$ is described in the Appendix. The power of the test can be obtained as

$$P_{\text{cap}}^{\mathcal{F}} = 1 - \{F_{\delta^+}(t_{4,\alpha/2}) - F_{\delta^+}(-t_{4,\alpha/2})\}\{F_{\delta^-}(t_{4,\alpha/2}) - F_{\delta^-}(-t_{4,\alpha/2})\}, \qquad (1.18)$$

where $F_{\delta^+}(\cdot)$ and $F_{\delta^-}(\cdot)$ denote the cumulative distribution function of a non-central t distribution with 4 degrees of freedom and non-centrality parameters $\delta^+ = (2\sum_{i \notin \mathcal{F}} \beta_i + 2\sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{F}} \beta_{ij})/(1.23\sigma)$ and $\delta^- = (2\sum_{i \notin \mathcal{F}} \beta_i - 2\sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{F}} \beta_{ij})/(1.23\sigma)$ respectively. Similar to the swapping phase, the power of the capping test is small when both $\sum_{i \notin \mathcal{F}} \beta_i$ and $\sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{F}} \beta_{ij}$ are small.

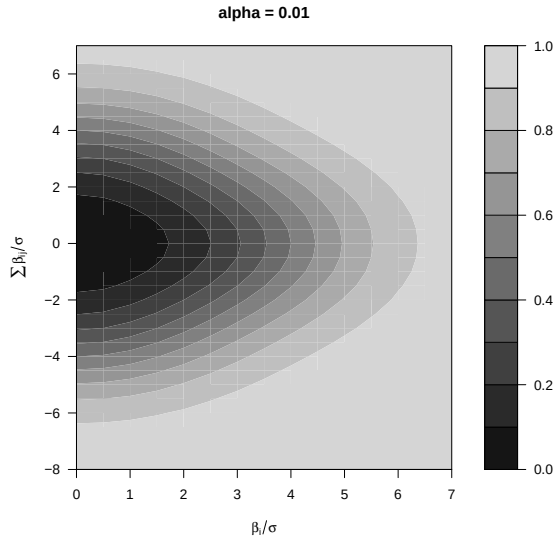


Figure 1.7: Power of Swapping to detect active factor x_i as a function of β_i/σ and $\sum_{j\neq i}\beta_{ij}/\sigma, \alpha = 0.01$.

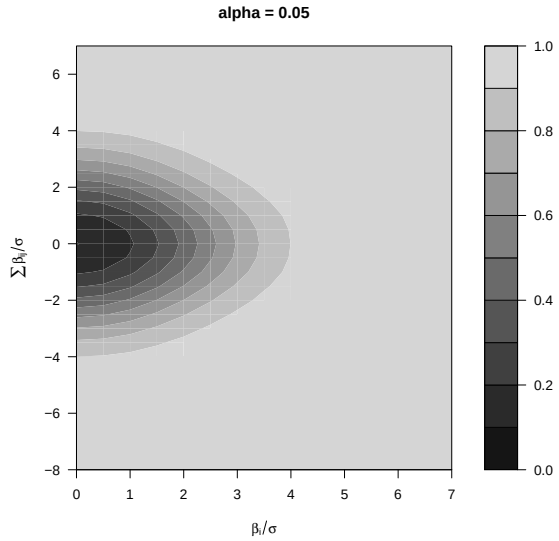


Figure 1.8: Power of Swapping to detect active factor x_i as a function of β_i/σ and $\sum_{j\neq i}\beta_{ij}/\sigma, \alpha = 0.05$.

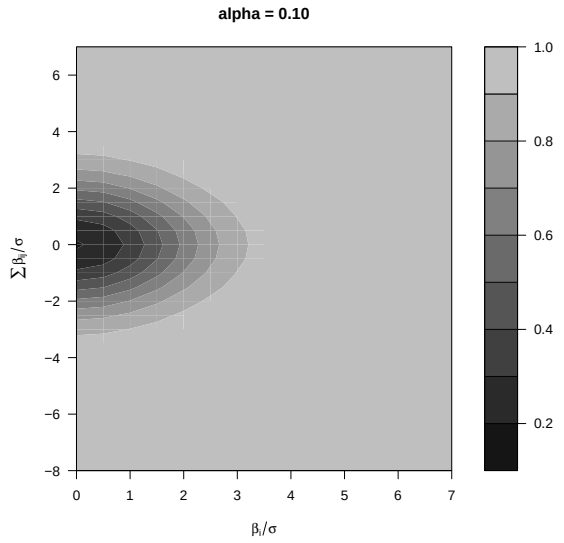


Figure 1.9: Power of Swapping to detect active factor x_i as a function of β_i/σ and $\sum_{j\neq i}\beta_{ij}/\sigma, \alpha = 0.10$.

1.4.2 The Overall Probability of Correct Screening

We shall now use the power functions developed in the previous three subsections to obtain the probability of correct screening of p ($< k$) active factors under model (1.12) for a specific order in which the k factors are investigated, assuming that the best and worst levels of each factor are correctly identified. Without loss of generality, let $x_1, \ldots, x_p$ represent the p active factors and $x_{p+1}, \ldots, x_k$ the $k-p$ inactive factors. Consider the following two extreme situations: (i) the p active factors are investigated first (assume that VS investigates k factors in the sequence $x_1, \ldots, x_p, x_{p+1}, \ldots, x_k$) and (ii) the $k-p$ inert factors are investigated first (assume that VS investigates k factors in the sequence $x_k, \ldots, x_{p+1}, x_p, \ldots, x_1$). Then the probability of correctly identifying $\{x_1, \ldots, x_p\}$ as the *only* active factors is given by

$$P_{CI} = \begin{cases} P_I(1-\alpha)(\prod_{i=1}^{p} P^i_{\text{swap}})(\prod_{i=2}^{p-1} P^{\mathcal{F}_i}_{\text{cap}})(1 - P^{\mathcal{F}_p}_{\text{cap}}), & \text{for situation (i)} \\ P_I(1-\alpha)^{k-p}(\prod_{i=1}^{p} P^i_{\text{swap}})(\prod_{i=2}^{p-1} P^{\mathcal{G}_i}_{\text{cap}})(1 - P^{\mathcal{G}_p}_{\text{cap}}), & \text{for situation (ii)} \end{cases} \tag{1.19}$$

where $\mathcal{F}_i = \{1, 2, \ldots, i\}$, $\mathcal{G}_i = \{p, p-1, \ldots, p-i+1\}$; and $P_I, P^i_{\text{swap}}, P^{\mathcal{F}}_{\text{cap}}$ are defined by (1.14), (1.15) and (1.18) respectively.

Note that this probability will depend on the order in which the active factors are investigated if their impacts on the response are different. Consider an example where seven factors $\{x_1, \ldots x_7\}$ are being investigated to identify three active factors $\{x_1, x_2, x_3\}$ that affect the response through the following model:

$$y = 0.8x_1 + 0.7x_2 + 0.8x_3 + 0.4x_1x_2 + 0.3x_2x_3 + 0.4x_1x_3 + \epsilon, \tag{1.20}$$

where $x_i = -1$ or $+1$ according as the worst or best level of factor i is used and $\epsilon \sim N(0, \sigma^2)$. Note that we do not include an intercept term in this model since the tests described in Section 4.1 are independent of the intercept term β_0 in model (1.12). The powers associated with each individual hypothesis test at different stages are computed using (1.14), (1.15) and (1.18) with $\sigma = 0.20$. Assuming a 5% level of significance for each test, the probability of detecting x_1, x_2 and x_3 as active and $x_4, \ldots, x_7$ as inert can be computed for situations (i) and (ii) using (1.19) as 0.8737 and 0.5818 respectively.

1.4.3 Inference in the Presence of Three-Factor Interactions

In the previous subsections, we have analyzed the VS technique where three-factor interaction effects were ignored. We now extend this to situations with three-factor

interaction effects. Consider the following third-order model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \sum_{i<j,l} \sum_{j<l} \sum_{l=1}^{k} \beta_{ijl} x_i x_j x_l + \varepsilon. \tag{1.21}$$

Due to the presence of three-factor interactions, the test statistic for testing hypotheses at different stages of the VS approach have slightly different sampling distributions as described in the Appendix. The test procedures, however, are identical to those derived earlier. It is worth noting that under model (1.21), negative values of three-factor interactions (3fi's) are likely to reduce the power of the tests at different stages. However, positive values of 3fi's associated with a particular effect will increase the power of its detection as an active factor.

1.5 Robustness of VS with respect to Noise Variation and Accuracy of Engineering Knowledge

In the earlier sections we have seen that for a given model, the performance of the VS design (with respect to the probability of correct screening) depends on the level of inherent noise variation σ^2 and the degree of correctness of the engineers' knowledge. We shall now examine the effects of these variables (here we refrain from using the term "factors" to avoid mix-up with the original experimental factors to be screened by VS) and their interactions on the overall probability of correct screening using the results derived in Section 1.4. In this study, we consider seven factors $1, 2, \ldots, 7$ to be investigated, out of which three factors $1, 2, 3$ are actually active. The response y depends on these three factors through the second-order model given in (1.20). We consider the following three input variables that are known to affect the performance of VS:

1. A: Incorrect engineering assumption about setting of a particular factor (in this case we consider factor 3 without loss of generality). This variable has two levels: the $+$ ($-$) level corresponds to a correct (incorrect) assumption, which

Table 1.3: Design matrix and percentage of correct screening

Run	Setting of factor 3 (A)	Order of investigation (B)	σ (C)	$100 \times P_{CI}$
1	Correct	1-2-3-4-5-6-7	0.10	90.25
2	Correct	1-2-3-4-5-6-7	0.30	49.14
3	Correct	1-3-4-5-2-6-7	0.10	73.6
4	Correct	1-3-4-5-2-6-7	0.30	43.5
5	Correct	7-6-5-4-3-2-1	0.10	59.87
6	Correct	7-6-5-4-3-2-1	0.30	33.96
7	Incorrect	1-2-3-4-5-6-7	0.10	90.23
8	Incorrect	1-2-3-4-5-6-7	0.30	20.24
9	Incorrect	1-3–4-5-2-6-7	0.10	70.1
10	Incorrect	1-3-4-5-2-6-7	0.30	13.5
11	Incorrect	7-6-5-4-3-2-1	0.10	59.86
12	Incorrect	7-6-5-4-3-2-1	0.30	9.63

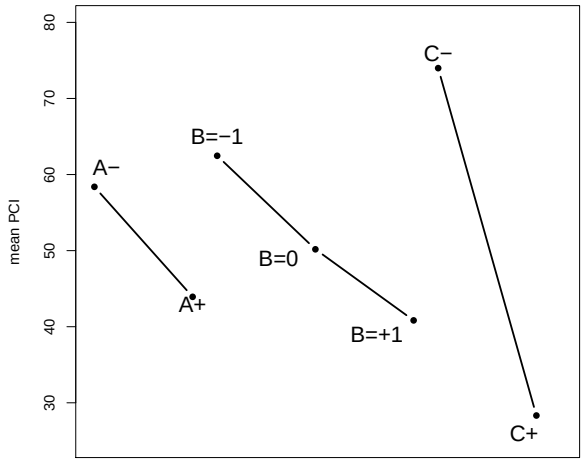
means that the coefficients of x_3, x_{23} and x_{13} in model (1.20) are 0.8 (-0.8), 0.3(-0.3) and 0.4(-0.4) respectively.

2. B: Incorrect engineering assumption about relative importance of factors 1-7. We consider three levels of this variable: the -1 level corresponds to the correct order 1-2-3-4-5-6-7, $+1$ level corresponds to the completely reverse (and incorrect) order 7-6-5-4-3-2-1, and 0 corresponds to 1-3-4-5-2-6-7.

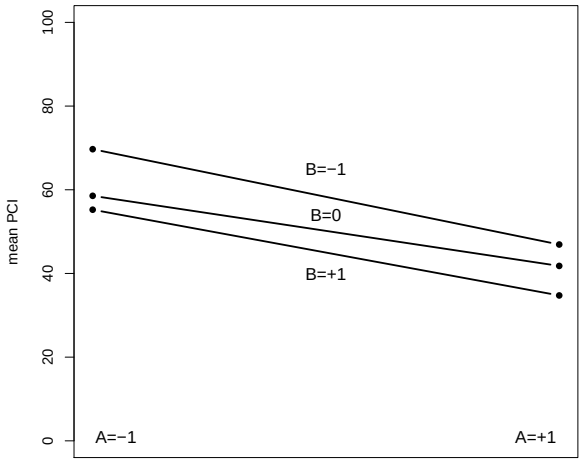
3. C: The standard deviation σ of the error term ϵ in model (1.20). The two levels of this variable are chosen as $\sigma = 0.10$ and $\sigma = 0.30$.

A full factorial experiment was designed with these three input variables, and for each combination, the percentage of correct screening ($100 \times P_{CI}$) was computed using (1.19). The results are summarized in Table 1.3.

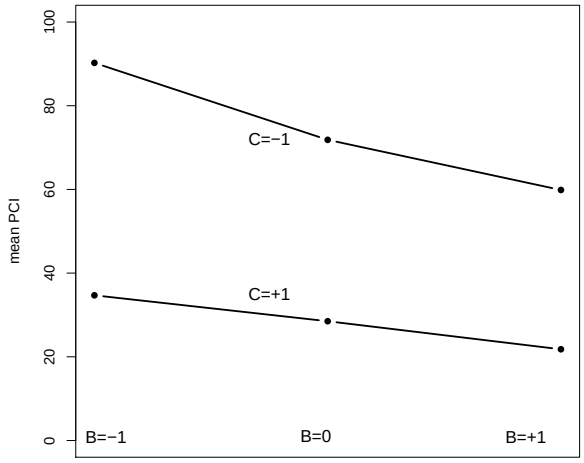
The data from Table 1.3 are summarized in the form of significant main effects and interaction plots (see Figure 1.10). All the three variables A, B and C are seen to affect the performance. As expected, the performance is poor with incorrect settings, wrong ordering or high error variance. In particular, a three-times increase in error variance is seen to have a strong effect on the performance. Two interactions $B \times C$ and $A \times C$ are also seen to affect P_{CI}. Clearly, higher noise level worsens the effect of



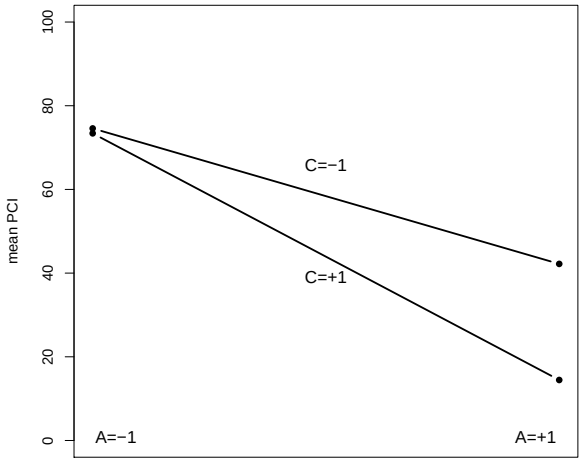
Main effects



$A \times B$ interaction



$B \times C$ interaction



$A \times C$ interaction

Figure 1.10: Plots of factorial effects from the designed experiment

lack of engineering knowledge on the performance. Also, the fact that a combination of incorrect setting, incorrect ordering and high noise (run 8 in Table 1.3) leads to a very low (9.63) percentage of correct screening, is indicative of the presence of a three-factor interaction $A \times B \times C$.

1.6 Concluding Remarks

We have investigated Shainin's Variable Search (VS) method with the objective of understanding it better, and also identifying the type of settings under which it does and doesn't work well. The results in Sections 1.3 - Section 1.5 have established that VS is a useful method for screening of factors if (a) the engineering assumptions about the directions of the factor effects on the response and the relative order of importance are correct and (b) the error variance is not very high relative to the main effects and sum of 2fi's involving each factor. The VS design permits unbiased estimation of the main effects under the assumption that interactions of order 3 and above are negligible. Further, it has some projection properties that permit independent estimation of main effects and 2fi's for a maximum of four active factors.

Thus, the VS method is likely to be particularly useful for screening active factors if the number of factors under investigation is large, e.g., $k \geq 15$ where it can lead to a significant saving of runs in comparison to a comparable fractional factorial design, particularly if higher order interactions are actually present in the model, or cannot be ruled out. In contrast, if the number of factors is not very large, e.g., $k \leq 10$, the experimenter's knowledge about the relative importance of factors is limited, and higher order interactions can be assumed away, fractional factorial designs or screening designs like Plackett-Burman designs will be a much better choice. Further, incorrect process knowledge and high error variance can result in poor performance of VS, both in terms of correctness of factor screening and run size.

1.7 References

Bhote, K. R. and Bhote, A. K. (2000). *World Class Quality*, Amacom: New York.

Box, G. E. P. and Tyssedal, J. (1996). "Projective Properties of Certain Orthogonal Arrays". *Biometrika* 83, pp. 950-955.

Box, G. E. P. and Hunter, J. (1961). "The 2^{k-p} Fractional Factorial Designs: Part I". *Technometrics* 3, pp. 311-351.

Ledolter, J. and Swersey, A. (1997). "Dorian Shainin's Variable Search Procedure: a Critical Assessment". *Journal of Quality Technology* 29, pp. 237-247.

Miller, A. and Sitter, R.R. (2001). "Using the Folded-Over 12-Run Plackett-Burman Design to Consider Interactions". *Technomterics* 43, pp. 44-55.

Plackett, R.L. and Burman, J.P. (1946). "The Design of Optimum Multifactorial Experiments". *Biometrika* 33, pp. 305-325.

Shainin, D. (1986). "Better than Taguchi Orthogonal Tables". *ASQC Quality Congress Transactions*, Anaheim, CA, pp. 446-481.

Shainin, D. and Shainin, P. (1988). "Better than Taguchi Orthogonal Tables". *Quality and Reliability Engineering International* 4, pp. 143-149.

Steiner, S. H.; Mackay, R. J.; and Ramberg, J. S. (2008). "An Overview of the Shainin System^{TM} for Quality Improvement". *Quality Engineering* 20, pp. 6-19.

Verma, A. K.; Srividya, A.; Mannikar, A. V.; Pankhawala, V. A; and Rathnaraj, K. J. (2004). "SHAININ Method: Edge Over Other DOE Techniques". *IEEE International Engineering Management Conference, Proceedings* 20, pp. 6-19.

Wu, C. F. J. and Hamada, M. (2009). *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York, Wiley, 2nd edition.

CHAPTER 2

VARIANCE MODELING FOR ROBUST PARAMATER OPTIMIZATION IN COMPUTER EXPERIMENTS

2.1 Physical Process, Stochastic Simulator and Statistical Emulator for A Nanoparticle Fabrication Process

Often, conducting physical experiments is costly or/and time consuming. For example, the growth of nanoparticles (Xu et al. (2009)) requires about a day for completion. Moreover, preparation and characterization of nanoparticles are expensive (Dasgupta et al. (2008)). Engineering knowledge presented in chemical kinetics or differential equations can be used to build simulators for exploring process behaviors. Stochastic components can be added to model certain uncertainties. The real-life application in this thesis uses a stochastic simulator to generate nanoparticle fabrication data.

Experiments on simulators are generally cheaper than physical experiments, but can take a few hours to complete a run, for a batch of nanoparticles. In such situations, an emulator based on statistical models (e.g., regression, kriging) is used to approximate the simulator. Though not as accurate as the simulator, it is faster than the actual simulator. It can be used for process optimization or managerial what-if analysis, where a rough knowledge about the variables on the process can be explored. A few physical experiments are later conducted to ascertain the findings from the analysis using the simulator and the emulator. If the results from the simulator/emulator do not match these results, then the simulator should be modified, and the process repeated. This process is illustrated in Figure 2.1.

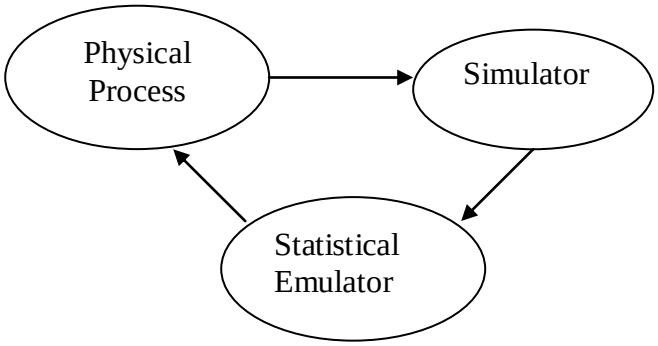


Figure 2.1: General framework for analyzing a process

In this research, our goal is to obtain the robust setting for a process. The simulator is used to obtain responses (e.g., average, standard deviation, load or yield from a batch of nanoparticles) for a limited number of design points. See Woo (2010) for the engineering background of a nanoparticle fabrication process.

2.2. Introduction

Computer simulations have gained popularity in many engineering fields. While investigating process behaviors, they are often preferred over physical experiments. The main reason for this is their advantage in terms of cost and time. As an example, consider production of nanomaterials. Characterization of nanomaterials requires expensive advanced equipment like transmission electron microscope (TEM). Data collection using TEM is time consuming. Thus, using process knowledge and /or relevant engineering equations, computer simulator approximates system's actual process behavior. Since the process will not be known completely, a few physical experiments are conducted to confirm findings from simulations.

Computer simulations have other advantages too. They can be used when physical experiments are unsafe (e.g., nuclear reactions). In physical experiments, noise variables are often hard to control whereas they can be set precisely in simulations. Expert opinion from experienced engineers can be used to improve the accuracy of simulation. Stochastic components can also be included in the simulation which helps in realizing process variations. In this research, it is assumed that effect of certain noise variables on the response is understood, and can be implemented into the simulator.

Computer simulations are often based on complex engineering models. Hence the design should provide information about the whole experimental region. The deterministic nature of the model implies that the response remains the same when replicated. Hence, for modeling mean process behavior, replicates do not add any value. As in Figure 2.2, SFD spreads the design points, and also avoids forming replicates. For modeling mean process behavior, it retains this property even when the design is projected over a subset of variables.

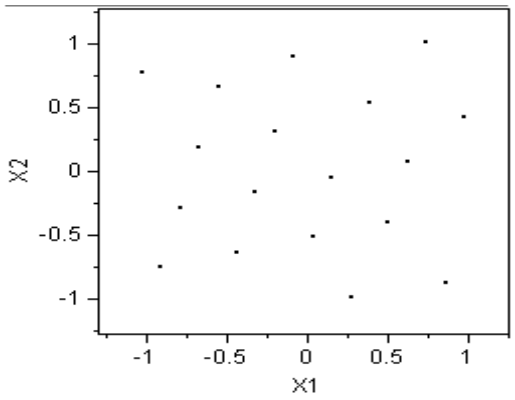


Figure 2.2: Space filling design

Taguchi (1986) introduced the concept of RPD. The goal of RPD is to obtain the setting of control variables such that the response is least sensitive to the variation in

noise variables. Not considering RPD in computer simulation can result in an "optimal setting" that is highly sensitive to the noise variables. Since the follow up physical experiments incur high cost and time, it is important to obtain the most accurate settings through computer simulations.

RPD in physical experiments has been researched extensively. Design for RPD can be broadly classified into two types, cross array and single array (Wu and Hamada (2009)). In cross array, each control variable setting is replicated for a set of noise variable settings. Thus, variance of response can be estimated for each control variable setting. Single array considers both control variables and noise variables together. Similarly, the analysis for RPD falls into two categories, namely variance derivation, and variance modeling. Variance derivation obtains variance expression as a function of the estimated mean model, e.g., Transmitted variance model (TVM) proposed by Shoemaker, Tsui and Wu (1991), uncertainty propagation (UP) proposed by Chen et al. (2006). Variance modeling considers mean and variance model separately in the "dual modeling" method procedure (Robinson et al. (2004)). Here, estimated variance is modeled in terms of the control variables to attain the robust setting.

Both single array and cross array have certain benefits. Single array is significantly smaller than the cross array, thereby demanding fewer resources. However, because of absence of replicates, it does not permit variance estimation and variance modeling. Cross array permits variance estimation, and hence permits both methods. Variance derivation is effective when the estimated mean model is accurate. However, it relies heavily on the estimated mean model implying that an incorrect or inaccurate model has a huge impact on the variance expression and hence the robust optimal setting

identified. Incorrect mean model impacts variance modeling too (Mcgrath and Lin (2001), Pan, G. (1999)). However its reliance on the mean model is less significant when compared to TVM and other variance derivation methods (Shoemaker et al. (1991), Chen et al. (2006)). TVM is applicable only to linear models. A more general formula for deriving the variance (Kang and Joseph (2009)), which is equivalent to TVM in case of linear models, is discussed below.

First, the mean system behavior (y) is modeled with respect to both control variables (**x**) and noise variables (**z**), followed by first order Taylor's series expansion around the nominal variables of noise (**z** = **0**), as follows.

$$y = f(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}) \approx f(\mathbf{x}, \mathbf{0}, \boldsymbol{\beta}) + \nabla_z f(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta})'_{z=0} \mathbf{z} + \varepsilon$$

where *f* represents the estimated model. The expression for variance is obtained from the above approximation as follows.

$$\text{var}(y) = \sigma_z^2 \nabla_z f(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta})'_{\mathbf{z}=\mathbf{0}} \nabla_z f(\mathbf{x}, \mathbf{0}, \boldsymbol{\beta})_{\mathbf{z}=\mathbf{0}} + \sigma_\varepsilon^2$$

where σ_z^2 and σ_ε^2 refer to the variance of noise and error part of the response, respectively. Note that if *f* is linear, then $f(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta})'$ depends only on the control variables (**x**), in a linear fashion. Thus variance can be obtained in terms of **x**, through control-noise interaction effects, which is equivalent to TVM.

The above procedure fails to identify the true variance model in certain situations. Its accuracy relies heavily on the accuracy of model *f* (Shoemaker and Tsui. (1993). Often, estimation of mean model is inaccurate when the true model is complex. This is supported by some examples discussed later, in Chapter 3. In some situations, even though *f* is accurate, it might fail to identify the true variance model. This is demonstrated through examples from a past research study, and a hypothetical example.

First, consider the study by Kunert et al. (2007) on certain metal forming experiments for robust parameter design. The experiments involved six control variables *A*, *B*,…, *F*, and three noise variables, m, n, and o. Both cross array (2^{6-3} x 2^{3-1}) and single array (2^{9-4}) designs were run separately. While dual modeling method can be applied on data from only cross array because of the presence of replicates, TVM is used to analyze data from both the designs because we consider a linear model. The authors analyzed the observed data to get the following results.

As part of the dual modeling procedure, mean model and variance model are identified separately using cross array data as follows:

Mean model: *F*, *E*, and *B* Variance Model: *A* and *D*

Next, all the design points were used for modeling the mean with respect to both control and noise variables. Seven effects were identified as follows:

F, *B*, *E*, *n*, *C*:*n*, *E*:*o*, and *A*:*B*:*n*.

TVM failed to identify the control variable *D* that affects the variance because of the absence of its interaction with noise variables. Similar result was observed when single array was analyzed. Note that the effects considered in the above models are aliased with one another. The authors argue TVM cannot identify the variance model, unless one assumes a mean model that involves large number of active effects, including three-factor-interactions that do not satisfy effect heredity. Such models can result in overfitting. The above example illustrates the case where the estimated mean model seemingly good, does not result in identification of true variance model. Such situations can arise even in cases where the relation between response and input variables is non-linear.

Next, consider the following example where even true mean model might lead to incorrect identification of variance model.

Example 1: Let the response (y) be related to x_1 (control variables) and z_1 (noise variable) as follows.

$$y = 3x_1 + 2z_1 + 3x_1z_1^2 \quad (2.1)$$

$\partial y/\partial z_1 = 2 + 6x_1z_1$ while $\partial^2 y/\partial z_1^2 = 6x_1$.

Here, using first order approximation, and evaluating it at $z_1 = 0$ (or when z_1 is small, close to 0) results in incorrect identification of effect of x_1 on the variance. Although such situations can be identified clearly in simpler case, it may not be trivial in case of non-linear modeling.

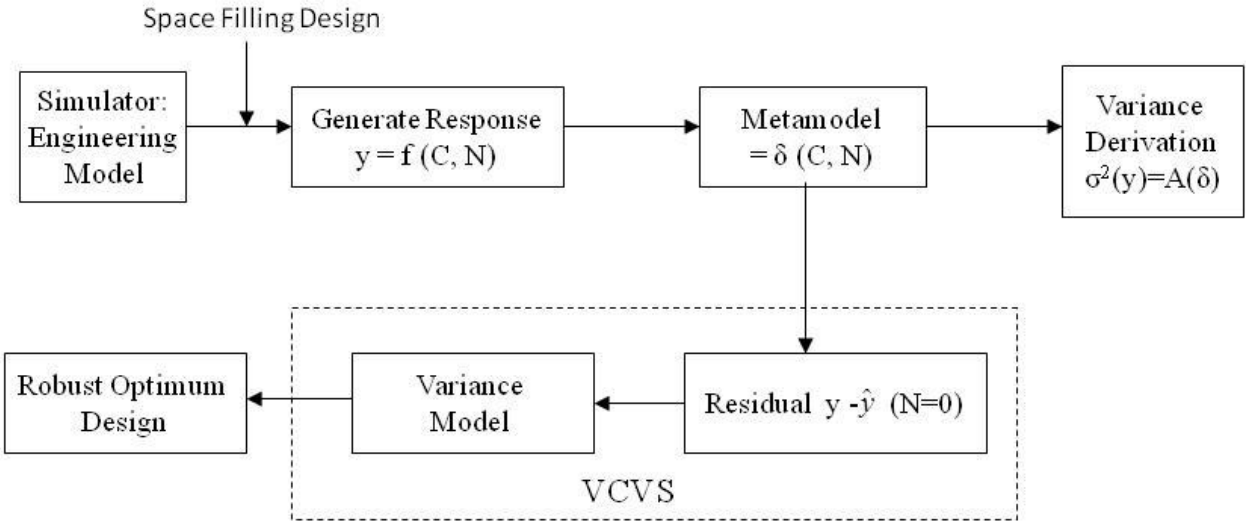


Figure 2.3: Schematic of the solution strategy

This research proposes variance modeling to support RPD study. It has benefits of both single array and variance modeling. As seen in Figure 2.3, simulator for the given system obtains responses for each setting of the design (SFD) comprising both control

and noise variables. Next, mean system behavior is modeled with respect to all variables. This step is common to both variance derivation and variance modeling methods. Variance derivation obtains variance as a function of the estimated response model. In contrast, the proposed method uses the residuals obtained through this model for variance modeling. The contribution of this research is towards modeling the variance in absence of natural replicates. Robust optimal setting can be attained by setting the mean closer to target and at the same time minimizing the variance, using the variance model obtained.

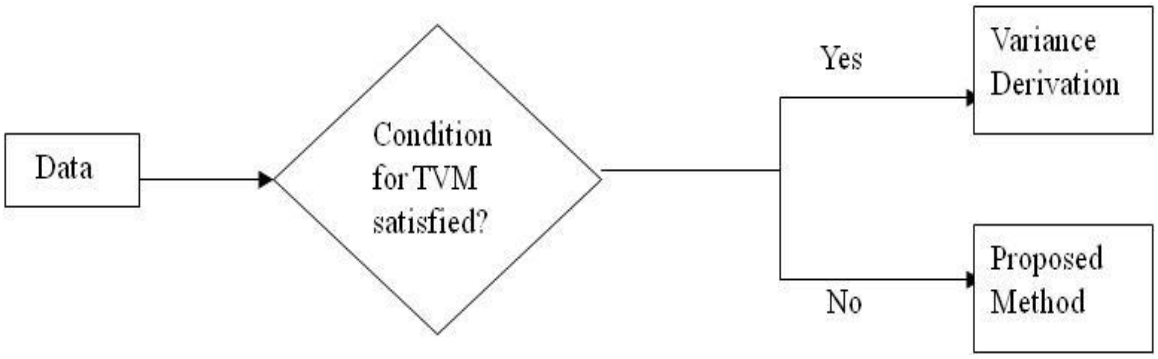


Figure 2.4: Choice of method for Robust Parameter Design

Both variance derivation and the proposed method have certain advantages. Thus, one should choose an appropriate method based on the observed data. Figure 2.4 depicts this comparison. This is achieved by comparing these methods under various conditions. Chapter 3 discusses this through some analytical and simulation studies.

The remainder of chapter 2 and chapter 3 is organized as follows. Section 2.3 discusses past research related to RPD in computer experiments. In Section 2.4, we propose a variance modeling method that makes use of variance-change-point (VC) method for grouping. Properties of this method in terms of model estimation are

discussed in Section 3.1. Next, impact of correlation on its performance is discussed in Section 3.2. Performance of the method is compared to other methods in Section 3.3. A real life example, nanoparticle growth process is used to demonstrate this method in Section 3.4. Finally concluding remarks are presented.

2.3. Literature Review

Since Taguchi (1986) introduced the concept of RPD, it has been researched extensively. Taguchi (1986) used cross array approach and proposed a two-step procedure to model the mean and dispersion for RPD studies in physical experiments. Box and Meyers (1986) and Montgomery (1990) use the projection properties of the factorial design to obtain robust design. Here, the design is collapsed onto only the significant variables results in replicates. In the above articles, variance estimation was possible because of the replicates. These approaches, classified as dual modeling, model both mean and variance separately. Cross array results in large design, especially if the number of control and / or noise variables is high. Hence single array consisting of both control and noise variables, was proposed. The design used is a factorial design with two levels for each variable. Linear models consisting of main effects and two-factor-interaction effects are used. Welch et al. (1990), Shoemaker et al. (1991) and Shoemaker and Tsui (1993) led the research analyzing this design wherein variance is derived from the estimated mean model. Vining and Myers (1990) used response surface methodology (RSM) approach for RPD where the estimated model can involve quadratic effects.

SFD for computer experiments does not have replicates. Unlike factorial designs, it does not possess projection properties that result in replicates. Thus variance modeling is a challenge. RPD for computer experiments has been researched to some extent. Bates

et al. (2005), Dellino et al. (2011), and Giovagnoli and Romano (2008) use the cross array approach to achieve robust setting. Bates et al. (2005) uses a SFD with both control and noise variables. Metamodel (e.g. Spline) fit based on observed data is used to predict the responses for a cross array. For each setting of the control array (SFD), 200 settings of noise variables are chosen randomly from their distributions. Dellino et al. (2011), and Giovagnoli and Romano (2008), in a similar approach, use cross array method for variance modeling. While the former method uses kriging model for prediction, the latter uses the simulation itself for obtaining the responses. The first two methods discussed above rely heavily on the estimated model. Though Giovagnoli and Romano (2008) does not depend on model estimation, it is time consuming because simulator should be run for all runs of the cross array, which can be very large.

Table 2.1: Literature survey of research related to robust parameter design

		With Replicates	Without Replicates
Physical Experiment	Variance Modeling	Taguchi (1986)	Montgomery (1990) Box and Meyers *(1986)*
	Variance Derivation		Shoemaker et al. (1991) Shoemaker and Tsui (1993)
Computer Experiment	Variance Modeling	Bates et al. (2005), Giovagnoli and Romano (2008), Dellino et al. (2011)	Adiga and Lu
	Variance Derivation		Dellino et al. (2010) Chen et al. (2006) Welch et al. (1990)

Some researchers (Dellino et al. (2010) and Chen et al. (2006)) derive variance from metamodels, using single array data. Both methods use SFD including both control and noise variables to obtain responses. Dellino et al. (2010) approximates the system behavior using polynomial mean model, and used RSM approach to obtain variance expression. In practice, the estimated model might be inaccurate because the true model in computer simulators are often complex. This method also relies heavily on presence on control-noise interaction effects. Chen et al. (2006) proposed UP to obtain the variance. The authors use tensor-product based basis functions to derive variance from mean model. Model structure is flexible, resulting in better mean model and hence better variance expression. However it still relies heavily on the estimated model. Table 2.1 depicts the literature survey in this field.

In this research, we propose a method to model variance based on single array data. The contributions of our research to literature are as follows. First, we use a new approach to group the responses to obtain pseudo replicates. Likelihood ratio test based change-point for variance ratio is used to determine the threshold for grouping. Hence, unlike most other methods, the group decisions depend on the observed response. To the best of our knowledge, this is the first time variance-change-point procedure is developed for locating a robust parameter setting. Next, the properties of the method are investigated in various conditions. In particular, it is analyzed when responses are correlated. In the past, effect of correlated response on variance modeling was considered in a factorial design setup with only two levels per variable. This research considers multiple levels for each variable and analyzes the performance when correlation structure is complex.

2.4. Solution Strategy

The proposed variance modeling method for RPD is illustrated in Figure 2.3. Response obtained from the simulator is modeled by an appropriate metamodel. Non-linear models with both control and noise variables are considered for estimating the true behavior because of the complex nature of computer simulations.

Denote the estimated model by $\hat{y} = \delta(\mathbf{x}, \mathbf{z})$, where ***x*** and ***z*** represent the vectors of control and noise variables respectively. Residuals (*r*) are computed as the difference between the actual response (from simulation), and the predicted value when noise variables are assumed at their nominal values ($\mathbf{z} = \mathbf{0}$).

$$r = y - \hat{y}(\mathbf{x}, \mathbf{z} = \mathbf{0}) = y - \delta(\mathbf{x}, \mathbf{0})$$

Here, δ is assumed to be a second order polynomial model. However, δ can assume any model form. This approach is particularly useful if the estimated model is interpolative. The traditional approach results in all residuals equals zero. The residuals obtained are used to model variance and obtain RPD solution. The variance model is assumed to be log-linear, of the following form.

$$\log(\sigma^2) = \mathbf{x}'\boldsymbol{\gamma} \qquad (2.2)$$

where $\boldsymbol{\gamma}$ is the vector of regression coefficients, and $\mathbf{x}$ is the vector of corresponding regressors and σ^2 is the corresponding variance . Henceforth, in this study, we assume variance model to be log-linear, as in (2.2).

Variance modeling requires replicates, which do not exist in SFD. The first task is to create pseudo replicate which groups a set of nearby data for variance modeling. This research assumes that nearby data have the same variance structure and the decision of whether a data point is grouped into a nearby cluster, is determined by VC procedure,

presented in Section 2.4.1. Next, the application of this procedure to estimate the variance model with multiple variables is illustrated. Increasing the number of variables considered for variance model results in higher number of groups thereby reducing the quality of variance estimate. Also, the variance estimate changes with the set of variables considered. Therefore, a variable selection procedure for variance modeling is formulated. Effectiveness of the procedure is analyzed through property investigation, discussing the parameter estimation in various situations, and the impact of correlated response on it. It is followed by a comparative study using hypothetical and real life examples.

2.4.1 VC Method

This section describes the procedure to determine the variance-change-point. Let x_i $i = 1,2,\ldots,n$, $x_i \le x_{i+1}$ be the values of predictor x, and y_i be corresponding response. x_i's are uniformly spread over the range [-1, 1]. The objective is to determine if the variable x affects the variance of y_i.

The effect of x on variance of y is analyzed by testing the following hypothesis.

$$H_0 : \sigma_j = s, \quad j = 1,2,\ldots,n. \tag{2.3}$$

$$H_1 : \sigma_j = s_{1,} \text{ if } x_j \le \tau$$

$$\sigma_j = s_{2,} \text{ if } x_j > \tau$$

where n is the size of design, τ is the variance-change-point i.e., the value of x, at which σ^2, the variance of y_i changes. τ is estimated by applying a well established Likelihood Ratio test (LRT) on the observed data. Each x_{ik} ($k = 1,2,\ldots,n$) is tested for variance-change-point through the following test statistic.

$$G_{k,n} = [(k-1)\ln(\frac{\hat{\sigma}^2}{\hat{\sigma}_1^2}) + (n-k-1)\ln(\frac{\hat{\sigma}^2}{\hat{\sigma}_2^2})]/C,$$

$$\text{where } C = 1 + (\frac{1}{k-1} + \frac{1}{n-k-1} - \frac{1}{n-2})/3$$

$\hat{\sigma}^2$ refers to the variance of all y_is, while $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the variance of y_is corresponding to x_i, before and the after the change-point (τ) respectively. In order to obtain a good variance estimate, k is chosen such that both groups of data consist of at least 20 observations, i.e., $20 \le k \le n-20$. The statistic $G = \max_k G_{k,n}$ is compared with the critical value $h_{n,\alpha}$ which is obtained through simulation studies. If $G \ge h_{n,\alpha}$ the corresponding value of x is the variance-change-point ($\hat{\tau}$). Else we declare τ does not exist, i.e., x does not affect the variance. The decision on the existence of change point depends on some other aspects like the variation of $G_{k,n}$, as seen in Section 2.4.2

Remarks:

(i) Hawkins and Zamba (2005) conducted a formal study of variance-change-point in a sequential experimental set up. By contrast, in this study the responses have already been obtained. Based on simulation studies, the authors obtained an approximated the critical values as follows ($n \ge 15$).

$$h_{n,\alpha} = -1.38 - 2.241\ln(\alpha) + [1.61 + 0.691\ln(\alpha)]/\sqrt{(n-9)} \text{ if } 0.001 \le \alpha \le 0.05$$

$$h_{n,\alpha} = 5 + 0.066\ln(n-9) \text{ if } \alpha = 0.05 \qquad (2.4)$$

(ii) Note that the $\hat{\tau}$ obtained from LRT has some uncertainty associated with it. $\hat{\tau}$ converges to its true value τ in probability, as $n \to \infty$ Chen et al. (2006a). i.e.,

$\hat{\tau} - \tau = O_p(1)$

In this research, it is assumed that $\hat{\tau} = \tau$. It is possible to induce the uncertainty into results by employing bootstrapping technique. We leave this to future research.

2.4.1.1. Variance Estimation for Variance Modeling

Sample variance given by (2.5) has been used in most studies to estimate variance.

$$s^2 = 1/(n-1)\sum_{i=1}^{n} r_i^2 \tag{2.5}$$

Variance model is then estimated by modeling the logarithm of this variance against the independent variables. However, this approach has certain drawbacks. Mcgrath and Lin (2000) state that presence of more than one active dispersion effect can result in identification of spurious effects in dispersion model. Brenneman and Nair (2001), and Nair and Pregibon (1988) reviewed methods used to estimate the variance, and the variance model. They argue that the variance model thus estimated is unbiased only under certain restrictive conditions. Note that sample variance (2.5) can be approximated as mean square deviation of the values from their mean value. Geometric mean of the squared deviation from the mean can be considered as an estimate for variance. Both the above studies discuss similar methods. Hence, the following statistic is proposed for estimating variance.

$$\tilde{s}^2 = \eta(\prod_{i=1}^{n} r_i^2)^{1/n} \tag{2.6}$$

Here, η = exp (1.27), a factor that makes the estimation unbiased. However, this statistic is sensitive if residuals are zero, or very small in magnitude. This problem can be resolved by adding a very small value, c $\neq$0 to all residuals.

More details are provided in Section 3.1. The application of VC method for identifying variance-change-point and subsequent variance modeling is illustrated using a simple example below.

Example 2: Consider the set of predictors $x_1, x_2, \ldots, x_6$ which might possibly influence the response *y*. Assume the true mean and variance models to be $\mu = 0$ and $\log(\sigma^2) = 5x_1$. Our objective is to estimate the mean model, and apply the above test to estimate the variance model.

A SFD involving all six variables is generated with run size $n = 80$. The range of the values of the variables is set to $[-1, 1]$. Responses are generated randomly based on the specified mean and variance model, and modeled against the variables $x_1, x_2, \ldots, x_6$. Since none of the variables were identified to affect the mean response, the residuals are the same as the response. VC method is applied on the responses as follows. For each variable x_i, $i = 1, 2, \ldots, 6$, the statistic G is computed and variance-change-point (τ_i) is estimated. The following discussion focuses on the procedure for variable x_1. The expression $G_{k,n}$ is computed for each value of k, $k = 20, 11, \ldots, n - 20$. Figure 2.5 shows that $G_{k,n}$ attains the maximum value of 169.4 at $x_1 = -0.01$ (i.e., $\tau_1 = -0.01$). This plot of $G_{k,n}$ is henceforth referred to as Generalized Likelihood Ratio (GLR) curve. Since G is much higher than the critical value $h_{n,\alpha} = 5.28$, the variable x_1 is said to affect the variance. The design is then split into two groups, namely hi ($x_1 \geq \tau_1$) and low ($x_1 < \tau_1$).

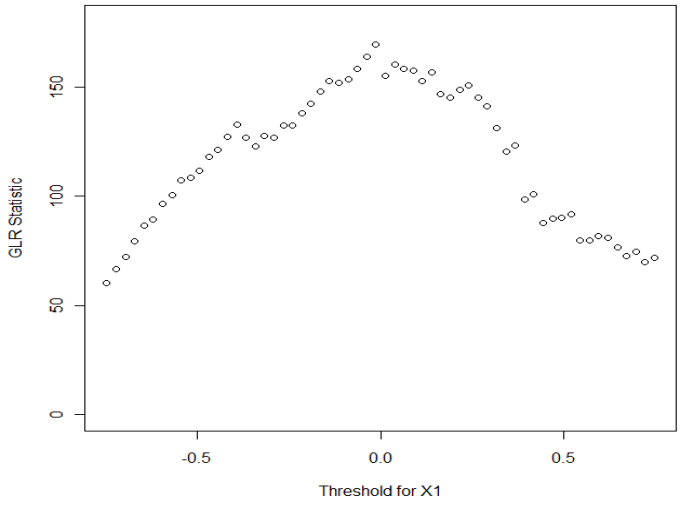


Figure 2.5: GLR Curve: example 1

The effect of the variable x_1 on the logarithmic variance γ_1 can be estimated as follows:

$$\gamma_1 = \frac{\log(\tilde{s}_{hi}^2) - \log(\tilde{s}_{low}^2)}{\mu_{hi} - \mu_{low}} = \frac{2.89 - (-2.31)}{0.532 - (-0.481)} = 5.14$$

where μ refers to the average x_1. Similarly, this procedure can be applied for each of the remaining variables.

2.4.2 Analysis of VC Method

This section explores VC method in various conditions, to improve its effectiveness. Factors such as design size, underlying variance model etc. are varied. This discussion is restricted to underlying variance models because it was found to be most important. The objective of this study is to facilitate the variable selection for variance modeling. First, one-variable case is considered, with design split into two groups. In addition to the LR test (LRT), we discuss the shape and other features of the GLR curve. Next, two-variable case is considered, where variance is modeled with respect to two variables. This method can be easily extended to the multi-variable case.

2.4.2.1 One-Variable Case

Here, VC method is applied with respect to x_1. However, the true variance model is not restricted to x_1. Various log-linear variance models are considered. It is assumed that mean model is accurately estimated. Thus the effect of inaccuracy in model estimation on identification of variance model can be ignored. Figure 2.6 illustrates GLR curves for few representative models using a SFD with 120 runs.

In the first case, $\log(\sigma^2) = 5x_1$, i.e., variance is affected only by x_1. The statistic G is significantly higher than the critical value $h_{n,\alpha}$ = 5.28. The GLR curve is smooth and bell shaped, with its peak near its center. In the second case, $\log(\sigma^2) = 5x_2$, the variance is not affected by x_1, but by x_2. The peak is much smaller than those for other models. $G_{k,n}$ is generally low, peaking only near the two extremes of the curve. Peaks are observed at extreme points, because one of the estimates σ_1 and σ_2 deviates from σ if variance estimation is inaccurate for small sample size. In comparison, in the first case, the effect of x_1 on variance dominates this effect. The third model is $\log(\sigma^2) = 5x_1 + 5x_1x_2$ where, apart from x_1, x_2 affects the variance through its interaction with x_1. Note that G is higher, compared to the first model because the effect of variables on the variance is higher. The GLR curve resembles the first case in shape, but is skewed to right. It also consists of small cluster of points parallel to the bigger one. This is an indication of the presence of an effect independent of the variable being considered, i.e., x_1. The fourth model has two main effects, x_1 and x_2. The GLR curve has many clusters of points that are parallel to each other. However, it still resembles a bell curve, with discontinuities in between. Such curves are also observed in case of $\log(\sigma^2) = 5x_2 + 5x_1x_2$, where a main effect other than the variable x_1 is active. Note that the G is much higher than the critical

value ($h_{n,\alpha}$ =5.31) in all the models except $\log(\sigma^2) = 5x_2$, where x_1 does not affect the variance.

To summarize the above discussion, if log-variance depends only on x_1 through a linear model, its effect can be clearly seen through the bell shaped GLR curve. Deviation from this shape implies that variance model is different. Skewed nature of the GLR curve indicates presence of interaction effects involving x_1. Parallel cluster of points indicates presence of another main effect. Randomness of the curve along with small G indicates that the variance is not influenced by x_1. In such situations, peaks are often observed near the extreme corners. Thus, investigating the GLR curves helps making an informed guess about possible variance model.

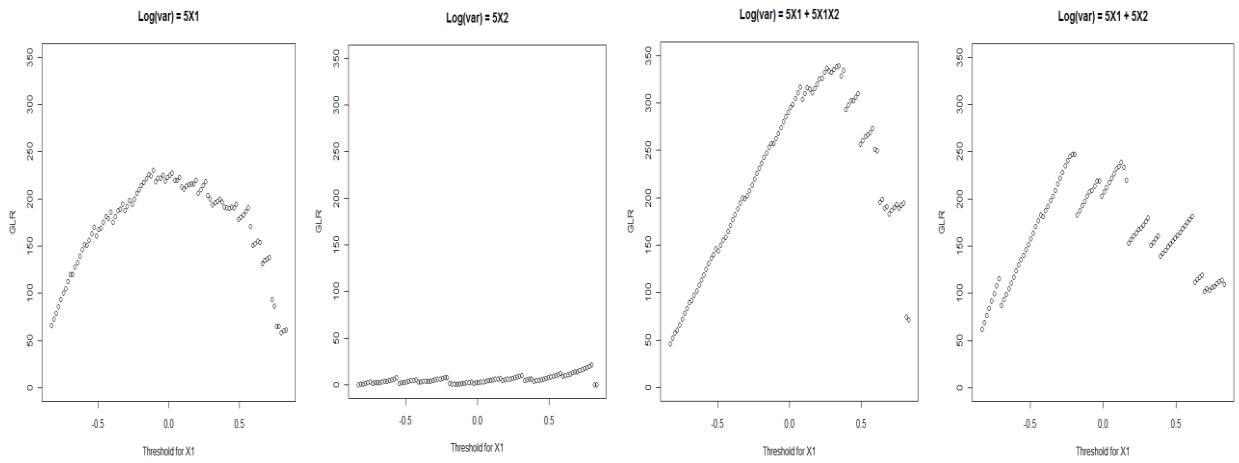


Figure 2.6: GLR curves for various log variance models

2.4.2.2 Two-Variable Case

Here, we consider estimation of variance model with respect to two or more variables. Consider two variables x_1 and x_2 that have potential effect on variance. The variance model is estimated as follows. First, variance-change-point for the two variables, $\hat{\tau}_1$ and $\hat{\tau}_2$ are determined. As shown in Figure 2.7, the design space is divided into four

parts, conditioned on $\hat{\tau}_1$ and $\hat{\tau}_2$. Variance is estimated for each region using pseudoreplicates, and modeled in terms of x_1, x_2, and their effects. This method can be extended to cases with more than two variables. In the two-variable case, testing significance of the effects is not direct because all three degrees of freedom are used for the main effects and the two-factor-interaction effect, and hence none for residuals. Solution for this problem is provided in the next subsection.

2.4.3 Variable Selection Procedure

Sections 2.4.1 and 2.4.2 discussed grouping and parameter estimation with respect to one or more variables. As the number of variables in the model increases, the number of groups increases at a faster rate resulting in fewer points in each group. Thus, it is important to include only the variables that impact the variance, for grouping. Moreover, the variance estimate obtained changes according to grouping. Identifying significance of model parameters requires sufficient degrees of freedom for the residuals of the regression model.

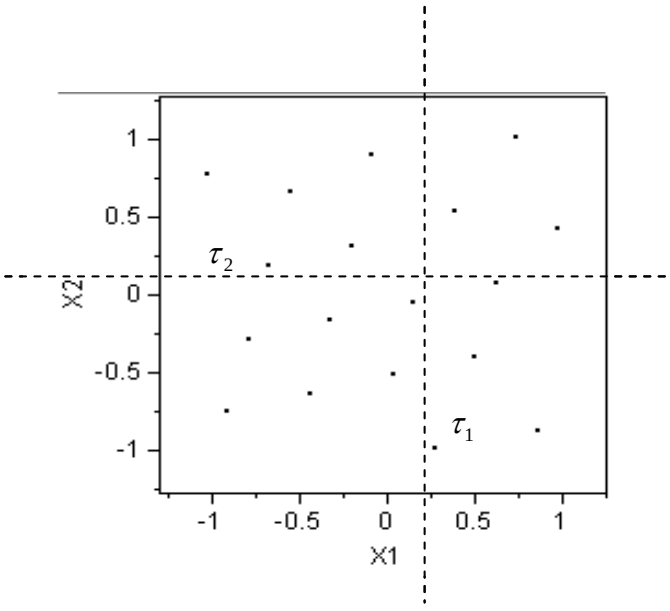


Figure 2.7: Splitting the design points into regions, based on the individual thresholds

Considering these, we propose a variable selection for modeling variance, as follows. Henceforth, in this study, this method is referred to as Variance-change-point based Variable Search method (VCVS).

Consider a system whose outcome is a function of p control variables $x_1, x_2, \cdots, x_p$ and p_N noise variables $z_1, z_2, \cdots, z_{p_N}$. Assume that residuals have been obtained based on an appropriate mean model. The three-step procedure for identifying the variables and the corresponding effects in the variance model is discussed below.

Step 1: Identification of variance-change-point.

(i) Apply VC method to each x_i to obtain G_i ($i = 1, 2, \cdots, p$). Higher G_i indicates the corresponding variable has higher impact on variance.

(ii) Analyze the shape and other features of the GLR curve (discussed in Section 2.4.1). Ignore the variables which are identified to have no effect on variance. Thus, only a subset $S \subseteq \{1, 2, \cdots, p\}$ of variables is considered in the next step.

Step 2: Preliminary Screening:

Each x_i, $i \in S$ is tested for its effect on variance through its main effect, or two-factor-interaction, as follows.

(i) For a given x_i, consider its main effect and two of its two-factor-interaction effects, say x_j, x_k $j \neq k$, $j, k \in S$. Apply stepwise regression to obtain the model.

(ii) Repeat (i) for all (j, k). Identify the best model M_i over all j, k, based on R^2_{adj} metric.

(iii) Repeat (i) and (ii) for all $x_i \; i \in S$ to obtain M_i. $i \in S$.

Step 3: Final model selection.

(i) Consider all the variables identified in Step1 and Step 2 through their main effects or two-factor-interactions. Let q ($q \leq p$) denote the number of such variables.

(ii) Divide the design into groups with respect to the above q variables Consider their main effects, and all two-factor-interactions having both components appearing among the main effects Use stepwise egression to obtain the final model. If $q \leq 2$, include 3-q inert variables for grouping, but not for modeling. Thus, we have additional degrees of freedom for the residuals that help in identification of significant effects.

Remarks:

1. This variable selection bears similarity with "Method I" (Hamada and Wu (1992)) for data with complex aliasing. Method I consists of an additional step where stepwise regression is applied, considering all effects satisfying effect heredity principle. Here, variables that are active only through their two-factor-interaction effects can be identified from the corresponding GLR curves.
2. Apart from the main effect, two two-interaction effects are considered, so that parameters of variance model can be tested for their significance. Fewer variables will result in insufficient degrees of freedom for residuals.

2.4.4 VCVS: Illustrative Examples

Here, VCVS is illustrated using few examples. Each example involves six control variables $x_1, x_2, \ldots, x_6$, and four noise variables z_1, z_2, z_3 and z_4 of which only few variables affect variance. A SFD with 80 runs is used to obtain the response. The objective is to identify and estimate the variance model, from observed data. Example 3 considers separate mean and variance models. Examples 4 and 5 consider complex mean models, and hence true variance model is not known. The accuracy in estimating mean model varies for each example. As seen later, estimation accuracy is a key factor that decides the effectiveness of this method.

Example 3: Consider the case with the true mean model and variance model defined as follows:

$$\mu = 0, \quad \log(\sigma^2) = 3.5x_1 + 3x_1x_4 \tag{2.7}$$

Responses are obtained from the normal distribution with mean and variance specified in (7). Residuals are obtained by fitting a polynomial regression model, which in this case is a constant function. Thus, residuals equal the response, i.e., $\hat{r} = y$.

Step 1: VC method is applied with respect to each control variable. Table 2.2 and Figure 2.8 show the results of these analyses. The GLR curve for x_5 has low peak occurring at the corner point. Hence, based on the discussion in Section 2.4.2, the variable x_5 can be ignored for further analysis. An overview of the GLR curves suggests that x_1 has the highest effect on the variance because of the highest peak, followed by x_4, x_2, x_6 and x_3 respectively. Also note that the GLR curve for x_1 is skewed towards right, indicating possibility of a two-factor-interaction effect involving x_1. GLR curves for x_2 and x_4 consist of parallel cluster of points, indicating presence of another main effect.

Step 2:

Consider three variables (x_1, x_2, x_3). The design is split into eight regions based on the thresholds for the corresponding variables. When the logarithmic variance for the groups is modeled against the effects (x_1, x_1x_2, and x_1 x_3), the following model is obtained. $\log(\sigma^2) = 2.89x_1$, with $R^2_{adj} = 0.72$. Similarly models for other combinations (x_1, x_2, x_4), (x_1, x_2, x_6), (x_1, x_3, x_4), (x_1, x_3, x_6), (x_1, x_4, x_6) are also identified. The best model (M_1) is identified (highest R^2_{adj} criterion) as

$$M_1: \log(\sigma^2) = 3.28x_1 + 2.7x_1x_4, \text{ with } R^2_{adj} = 0.88.$$

Similarly M_i are obtained for all x_i. None of the effects are identified for variables x_2, x_3, x_4 and x_6, i.e.,

$$M_2 = M_3 = M_4 = M_6: \log(\sigma^2) = \text{Constant}.$$

Thus x_1 and x_1x_4 are the only effects identified in step 1 and step 2. Thus, two variables, x_1 and x_4 are considered for step 3.

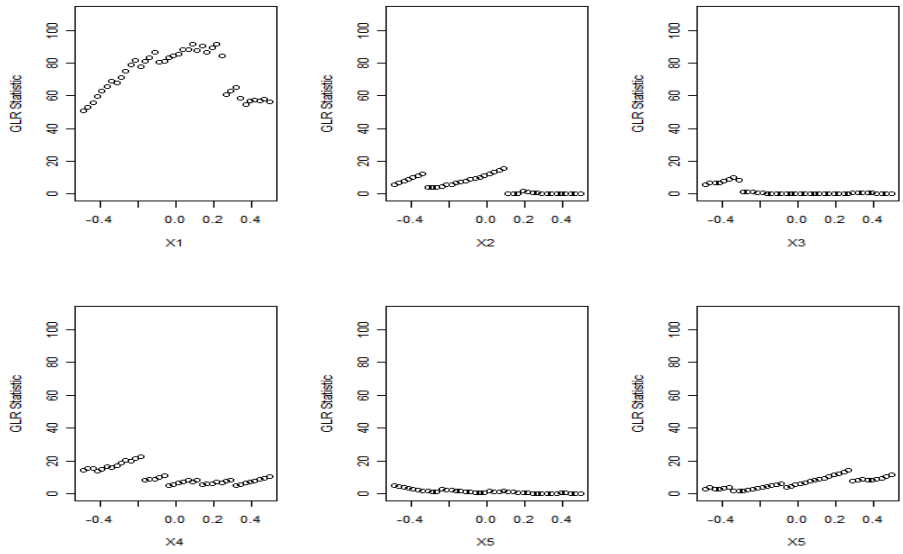


Figure 2.8: GLR curves corresponding to example 3

Table 2.2: Change-point for each variable, example 3

Variable	x_1	x_2	x_3	x_4	x_5	x_6
G	91.5	15.7	9.9	22.2	4.8	14.2
T	0.21	0.09	-0.34	-0.19	-0.49	0.26

Step 3:

Since the number of variables chosen $q = 2 < 3$, include one inert variable, say x_3, for grouping with τ_3=0. Stepwise regression is applied considering main effects and two-factor- interaction effects involving x_1, and x_4, to obtain the final model as follows.

$$\log(\sigma^2) = 3.7x_1 + 2.8x_1x_4, \text{ with } R^2_{adj} = 0.89.$$

VCVS has successfully identified the variance model, and estimated the model parameters with reasonable accuracy.

Example 3 considered a situation where the true variance model is known exactly. Next, in examples 4 and 5, we consider situations where true variance models are not known. Mean behavior is assumed to be non linear in terms of control and noise variables. As discussed before, it is approximated by a second order polynomial regression model. If the model is non-linear or unknown, true variance cannot be obtained analytically.

Example 4: Consider a hypothetical system where the response is related to predictors as follows. $y = \exp(1.2x_2 + z_2) + \exp(1.2x_2 + z_1)$. The mean model can be approximated by the following second order polynomial regression model.

$$\hat{y} = 2 + 3.2x_2 + 1.77x_2^2 + 1.4z_1 + 0.5z_1^2 + 1.3z_2 + 0.55z_2^2 + 1.4x_2z_1 + 1.3x_2z_2 + 0.28x_1x_2$$

with R^2_{adj} = 0.986. Hence it is a very good approximation of the true mean model.

The residuals are obtained as $\hat{r} = y - (2 + 3.2x_2 + 1.77x_2^2 + 0.28x_1x_2)$. Variance model is identified as follows.

Step 1:

VC method is applied to control variables x_1 through x_6. The corresponding test statistic and the GLR curves are provided in Table 2.3 and Figure 2.9 respectively. The effect of the variable x_2 on the variance is very clear from the high G value. G_i for all variables except x_2 are lower than the critical value $h_{n,\alpha}$ =5.13 Hence x_i, $i \neq 2$ do not affect the variance, and can be ignored for further analysis.

Table 2.3: Change-point for each variable, example 4

Variable	x_1	x_2	x_3	x_4	x_5	x_6
G	1.01	44.82	1.36	1.11	1.65	1.22
T	0.42	0.04	0.16	0.49	0.16	-0.29

Step 2:

This step is redundant, since only one variable, x_2 is considered for analysis. Hence, skip to step 3.

Step 3:

Since only one variable has been chosen in the first and second steps, two more inert variables (say, x_1 and x_3) are included for grouping. Variance estimates are obtained for the eight groups, and modeled against x_1 to obtain the following variance model.

$\log(\sigma^2) = 0.4 + 2.17x_2$, with R^2_{adj} = 0.95.

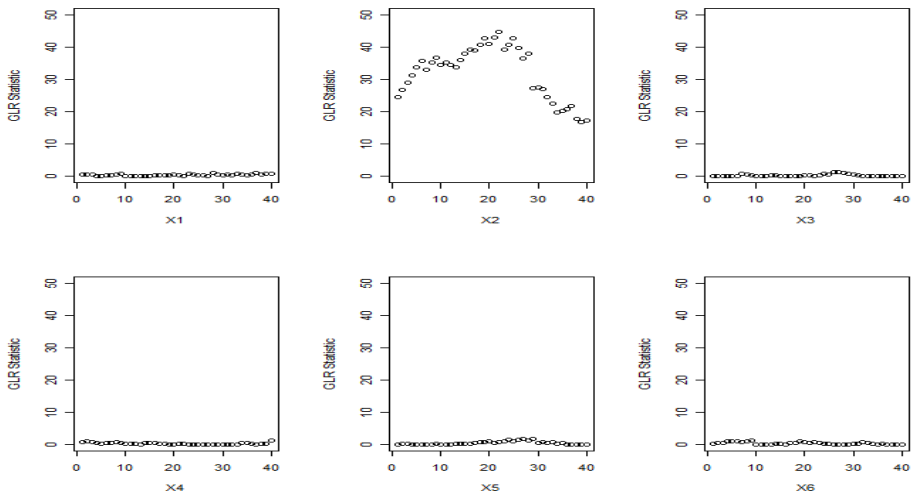


Figure 2.9: GLR curves for example 4

Example 5:

Consider a system where the response (y) is related to predictor variables through $y = \exp(x_1 + 1.2x_2 + z_1)/(x_1 + 0.5z_2)$. The following polynomial model is fit to observed data.

$$\hat{y} = 6.4 - 11.4x_5^2 - 11.4z_2^2$$

with R^2_{adj} =0.10 implying model estimate is inaccurate. Note that in practice, the true mean model is not known. Residual $\hat{r}$ is obtained as

$$\hat{r} = \hat{y} - (6.4 - 11.4x_5^2).$$

Variance model is identified as follows.

Step 1: VC method is applied to control variables x_1 through x_6. The corresponding results and plots are provided in Table 2.4 and Figure 2.10 respectively. x_1 appears to affect the variance most, followed by x_5. It is easy to see that x_3 does not affect variance (low G with peak observed at the extreme points), and hence excluded from further analysis.

Table 2.4: Change-point for each variable, example 5

Variable	x_1	x_2	x_3	x_4	x_5	x_6
G	42.85	9.42	14.41	9.05	21.27	10.35
T	-0.47	0.24	-0.47	0.49	0.09	-0.44

Step 2:

The best model obtained by comparing models with main effect x_1, and two of its two-factor- interactions is as follows:

$$M_1:\ \log(\sigma^2) = -3.6x_1 \text{ with } R^2_{adj} = 0.63.$$

Following similar approach, we obtain the following result:

$$M_2 = M_4 = M_6:\ \log(\sigma^2) = \text{constant}.$$

$$M_5:\ \log(\sigma^2) = -2.1x_5, \text{ with } R^2_{adj} = 0.40.$$

Hence, x_1 and x_5 have been identified in the first two steps.

Step 3:

Stepwise regression is applied considering the main effect and two-factor-interactions involving x_1 and x_5. Since only two variables have been identified, consider one more variable, say x_6 for grouping with $\tau_6 = 0$. The final model is obtained as follows:

$$\log(\sigma^2) = -3.45x_1 \text{ with } R^2_{adj} = 0.78.$$

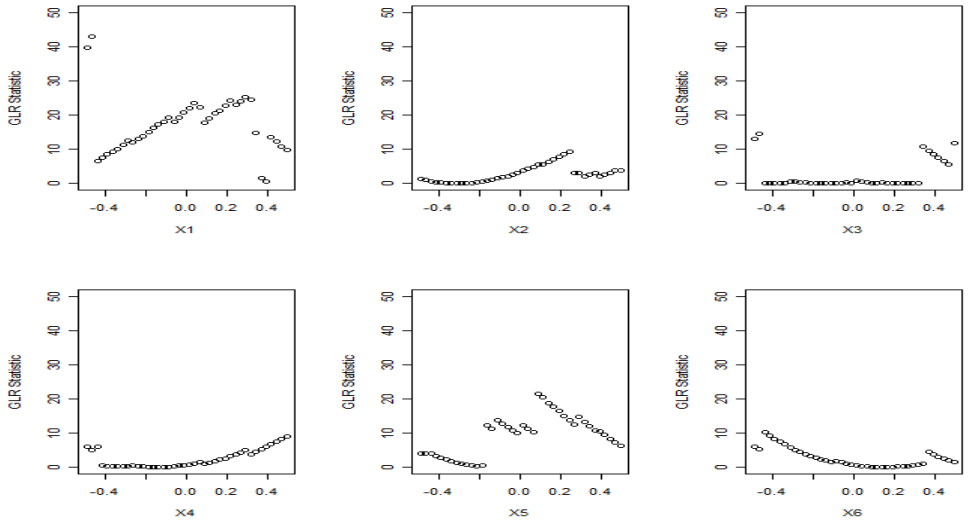


Figure 2.10: GLR curves for example 5

Contribution of this Chapter to Research:

- Develops a procedure to model the variance for RPD, in absence of replicates.
- Propose grouping the design points in order to obtain pseudo replicates. In particular, this research uses variance change point approach for grouping
- Formulate a three step variable selection procedure to identify and estimate the variance model.

CHAPTER 3

PROPERTY INVESTIGATION

This section analyzes properties of VCVS. First, we analyze parameter estimation in terms of unbiasedness and standard error of the estimates. Next, the impact of correlated response on variance modeling is discussed. Empirical results are obtained through a simulation study to support this. The proposed VCVS method is then compared to other existing methods. Finally, it is applied to a real life example, nanoparticle fabrication process.

3.1 Estimation of Variance Model

As discussed in (2.6), variance for each group is estimated as the geometric mean of residual squares, i.e., $\tilde{s}^2 = \eta(\prod_{i=1}^{n_g} r_i^2)^{1/n_g}$. Variance model is obtained by modeling log-variance against the control variables, as in (2.2). Here, variance estimation is discussed followed by estimation of model parameters.

3.1.1 Estimation of Variance

Sample Variance is the minimum variance unbiased estimator (MVUE) of population variance. Hence, the proposed estimate for variance $\tilde{s}^2$ is not the best estimator of variance. However, this statistic leads to better identification of log-linear variance model. Mcgrath and Lin (2001) used similar statistic to avoid spurious effects being identified. However, if some observations equal zero, $\tilde{s}^2$ goes to infinity. One can add a very small number to avoid this problem.

Let $\mu_1,\ i = 1, 2, \cdots, \kappa$ denote mean value of design variable(s) *x*, for group *i*. Here, κ denotes the number of groups that the design is divided into. Let σ_i^2 and $\tilde{s}_i^2$ represent the corresponding true variance and its estimate. The variance model is obtained by modeling $\log(\tilde{s}_i^2)$ against μ_i. Hence, we analyze the properties of estimates of $\log(\tilde{s}_i^2)$ instead of $\tilde{s}_i^2$. The following results illustrate the basic properties of this estimate.

Proposition 1. The logarithm of the proposed statistic $\log(\tilde{s}_i^2)$ is an unbiased estimator of $\log(\sigma_i^2)$

Proof: See Appendix

Variance of the above statistic is obtained as follows:

$$\text{var}(\log(\tilde{s}^2) = \text{var}(\frac{1}{n_g}\sum_{i\in\phi_g}^{n_g}\log(r_i^2)) = \frac{1}{n_g^2}\text{var}(\sum_{i\in\phi_g}^{n_g}\log(r_i^2)) \tag{3.2}$$

where ϕ_g denotes the set of points in group *g* and n_g is the number of points in ϕ_g. It is difficult to obtain variance of the statistic (3.2) analytically. Empirically, once can estimate the above variance through a very large sample of random variables from its distribution, as follows:

$$v_0 = \text{var}(\log(r_i^2)) \approx 4.934$$

Thus, increasing n_g decreases the variance in (3.2).

3.1.2 Parameter Estimation

Variance model (2.2) is fit using the variance estimates for each group. Here, we discuss the estimation of parameters of the variance model, in terms of unbiasedness and standard error of the estimate. It is assumed that observed responses are uncorrelated.

Let $r_1, r_2, \ldots, r_{n_g}$ be the residuals corresponding to the design points in group *g*. Estimate for log-variance for group *g*, i.e., $\log(\sigma_g^2)$ is given by

$$\frac{1}{n_g}\sum_{i\in\phi_g}^{n_g}\log(r_i^2)$$

Consider the simple case of two groups with respect to one variable say *x*, with n_1 and n_2 points respectively. Then, the estimate for the coefficient of *x*, in the variance model is given by

$$\hat{\gamma}_1 = \frac{\frac{1}{n_2}\sum_{i\in\phi_g, i=1}^{n_2}\log(r_i^2) - \frac{1}{n_1}\sum_{i\in\phi_g, i=1}^{n_1}\log(r_i^2)}{\mu_2 - \mu_1} \tag{3.3}$$

where μ_1 and μ_2 denote the mean value of *x* for groups 1 and 2 respectively.

Proposition 2: The estimator of model parameter, given by (3.3) is unbiased.

Proof:

This result can be extended to cases involving multiple variables, and higher number of groups. See appendix for details.

Next, consider the variance of the above estimate. Variance of the estimate of model parameter in (3.3) can be expressed as follows:

$$\begin{aligned}\mathrm{var}(\hat{\gamma}_1) &= \mathrm{var}([\frac{1}{n_2}\sum_{i\in\phi_g, i=1}^{n_2}\log(r_i^2) - \frac{1}{n_1}\sum_{i\in\phi_g, i=1}^{n_1}\log(r_i^2)]/(\mu_2-\mu_1)) \\ &= v_0(\mu_2-\mu_1)^{-2}(1/n_1 + 1/n_2)\end{aligned} \tag{3.4}$$

The variance in (3.4) is a function of the variance of the individual terms. In multi-variable case with κ groups, variance-covariance matrix of parameters ($\hat{\boldsymbol{\gamma}}$) is given by

$$\Sigma = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2_{\tilde{s}\lg} \tag{3.5}$$

where $\Sigma = \text{var}(\hat{\boldsymbol{\gamma}})$ is the variance-covariance matrix, $\mathbf{X}$ is the design matrix with κ rows corresponding to each group, and the columns refer to each effect estimated by $\hat{\boldsymbol{\gamma}}$. $\hat{\sigma}^2_{\tilde{s}\lg}$ denotes the estimate of the variance of estimates $\log(\tilde{s}^2)$, which in turn, are the response for variance modeling, i.e.,

$$\hat{\sigma}^2_{\tilde{s}\lg} = \text{var}(\log(\tilde{s}_1^2), \log(\tilde{s}_2^2), \cdots, \log(\tilde{s}_\kappa^2))$$

The results are illustrated using an empirical study considering a few variance models.

Consider the following cases (i) $\mu = 0, \log(\sigma^2) = 5x_1$ (ii) $\mu = 0, \log(\sigma^2) = 5x_1 + 5x_2$ (iii) $\mu = 0, \log(\sigma^2) = 5x_1 + 5x_1x_2$ and (iv) $\mu = 0, \log(\sigma^2) = 2x_1$. Each case assumes zero mean but variance models are different. VCVS is applied to estimate variance model, using a SFD with 80 runs. The design space is divided into eight groups, based on the thresholds for three of the variables. In addition to x_1 and x_2, include an additional variable with $\tau = 0$. Variance is modeled with respect to the variables x_1, x_2 and their interaction effect.

The above steps are repeated is repeated 200 times, each time generating responses from the specified distribution. The summary of this study is illustrated in Figures 3.1 through 3.4. Distributions of the model parameters provided in each case confirm that the estimation is unbiased. Standard errors of estimates of the interaction effects are higher than that of the main effects and the intercept because the coefficient of the interaction effects in $(\mathbf{X}'\mathbf{X})$ is lower than those of the main effects, as all settings are in the range [-1, 1]. Standard errors also vary depending on presence or absence of other

active effects, to some extent. It is the lower when only one variable is active (cases (i) and (iv)), and slightly larger in presence of other active effects (cases (ii) and (iii)).

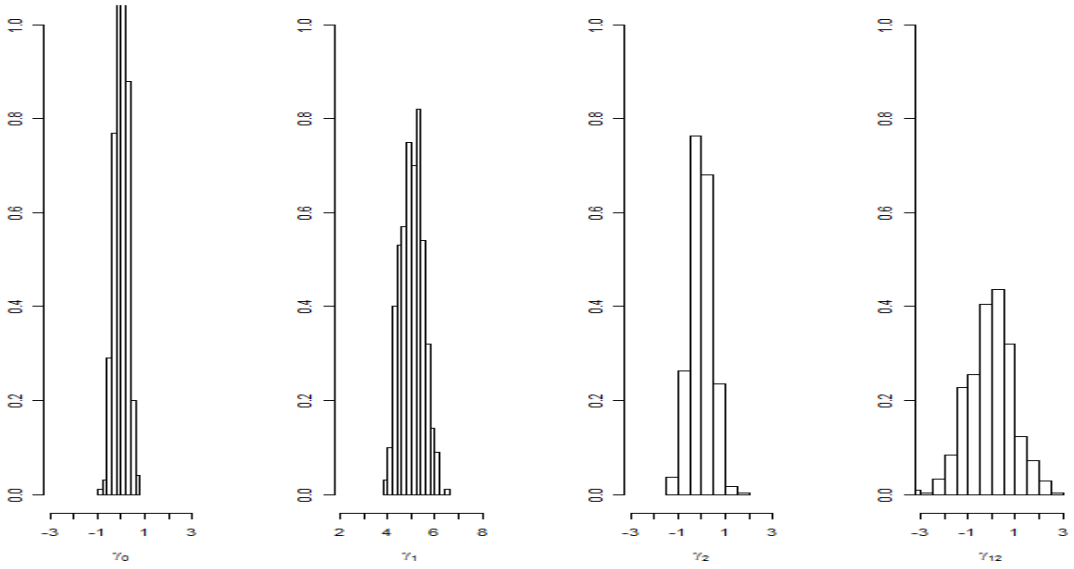


Figure 3.1: Distribution of parameter estimates: True variance model: $\log(\sigma^2) = 5x_1$

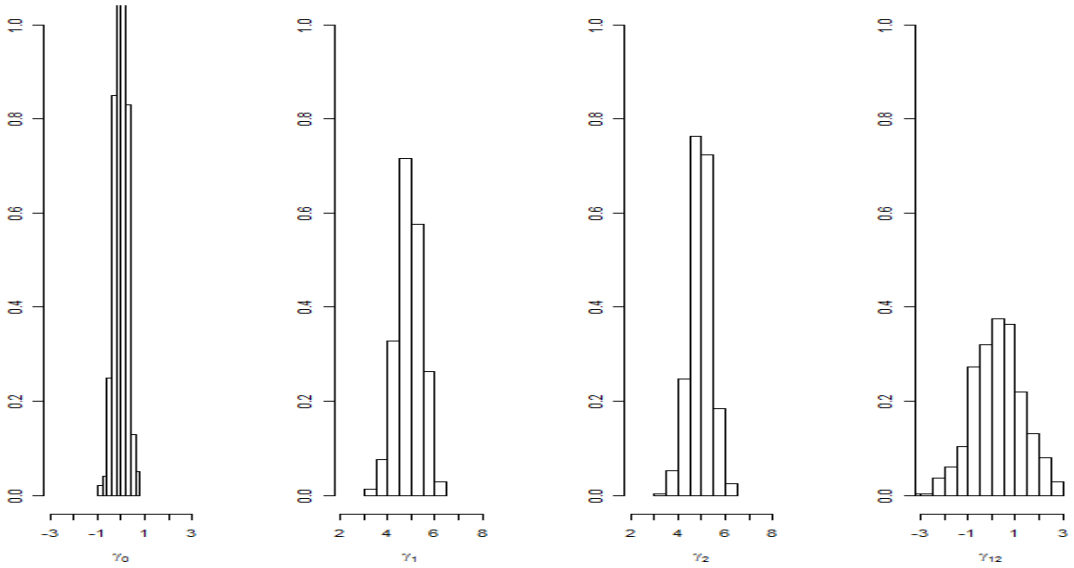


Figure 3.2: Distribution of parameter estimates: True variance model: $\log(\sigma^2) = 5x_1+5x_2$

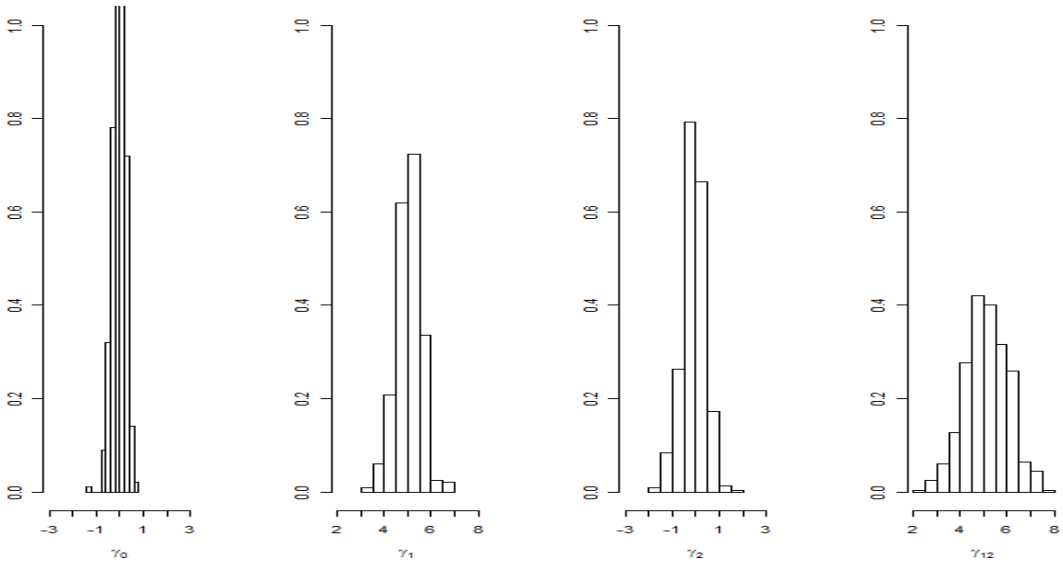


Figure 3.3: Distribution of parameter estimates: True variance model: $\log(\sigma^2) = 5x_1 + 5x_1x_2$

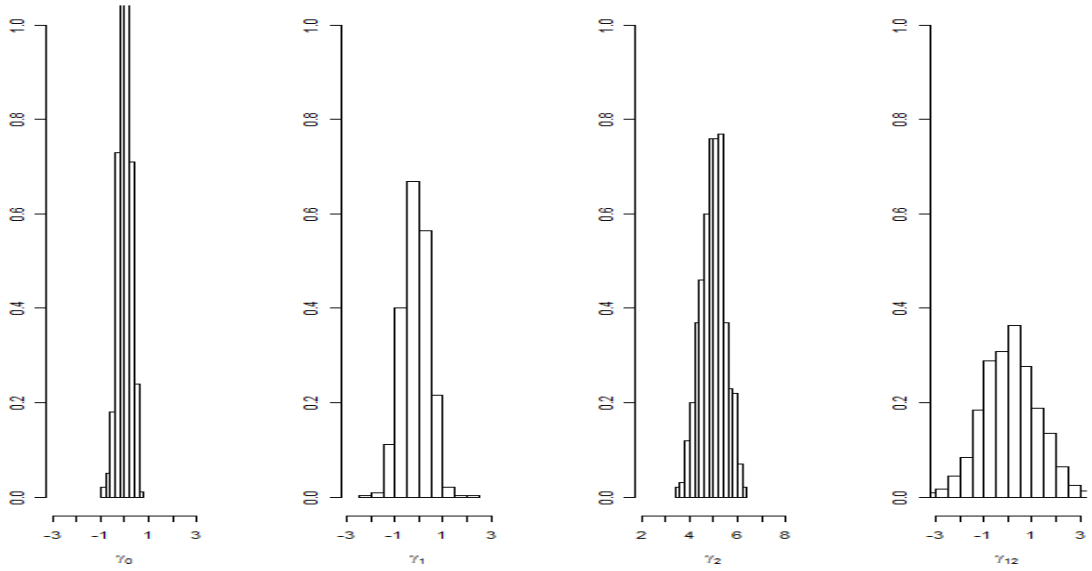


Figure 3.4: Distribution of parameter estimates: True variance model: $\log(\sigma^2) = 5x_2$

3.2 Correlated Response

In Section 3.1, responses (i.e., residuals) were assumed to be uncorrelated. This assumption is true in some situations. For example, the linear regression model assumes the residuals to be independently distributed. However, in practice, the observed responses are often correlated, e.g. spatial data. In such situations, metamodels such as kriging can be employed for modeling. Kriging models the inherent correlation structure in the observed data, to obtain a prediction model. Here, it is assumed that correlation between the responses cannot be ignored.

Correlation between the responses has been discussed extensively in literature. Most research has focused on modeling and predicting the mean behavior, based on data from computer experiments. Sacks et al. (1989), Santner et al. (2003), Fang et al. (2006), Cressie (1993) etc. discuss the modeling of correlated data from computer experiments and other correlated data. Joseph (2006), Joseph and Hung (2008), etc. discuss kriging methodologies to model correlated data. In case of linear models for mean modeling, the correlation affects the standard error of the estimates. Depending on whether the correlation is positive or negative, the standard error of the parameters are underestimated, or overestimated. Thus, correlation between responses can lead to identifying spurious effects, or missing some active effects.

Correlated response impacts variance estimation too. The impact of correlation on estimation of variance model and robust parameter design has not been studied extensively. Bergman and Hynen (1997) considers the correlation between responses in a factorial design set up dealing with only two levels for each variable. In contrast, in computer experiments, each variable is varied over multiple levels. Also, it allows more

flexibility in the correlation structure. Here, we analyze the impact of correlation on variance modeling in computer experiments. We assume non-negative correlation between responses, as it is intuitive. Moreover, negative correlation results in the correlation matrix not being positive semi definite matrix. First, we focus on identifying the effect of correlation on the results analytically. Later, an empirical study is performed to investigate the results.

3.2.1 Impact on Variance Estimation

Here, the impact of correlation on the estimation of log-variance is analyzed. Later, its effect on the parameter estimation for variance model is investigated. The findings are complemented by subsequent empirical study.

The estimate of log-variance is unbiased even in presence of the correlation. In Proposition 1 and its proof, it was seen that $\log(r_i^2) = \log(\sigma_i^2) + \log(\chi_1^2)$. Correlation between responses does not alter the expectation $E(\log(\chi_1^2))$ indicating that it does not variance estimation is unbiased, even in presence of correlation.

Next consider the variance estimate in (3.2). If the responses are correlated, then this variance can be expressed as follows:

$$\begin{aligned}
&\operatorname{var}(\log(\tilde{s}^2)) = \frac{1}{n_g^2}\operatorname{var}\Big(\sum_{i\in\phi_g, i=1}^{n_g}\log(r_i^2)\Big) \\
&= \frac{1}{n_g^2}\Big(\sum_{i\in\phi_g, i=1}^{n_g}\operatorname{var}(\log(r_i^2)) + 2\sum_{i\in\phi_g, i=1}^{n_g}\sum_{j\in\phi_g, j<i}\operatorname{cov}(\log(r_i^2),\log(r_j^2))\Big) \\
&= \frac{1}{n_g^2}\Big(\sum_{i\in\phi_g, i=1}^{n_g}\operatorname{var}(\log(r_i^2)) + 2\sum_{i\in\phi_g, i=1}^{n_g}\sum_{j\in\phi_g, j<i}\rho_{ij}^{*}\sqrt{\operatorname{var}(\log(r_i^2))\operatorname{var}(\log(r_j^2))}\Big)
\end{aligned} \tag{3.6}$$

$$= \frac{1}{n_g^2} \sum_{i \in \phi_g, i=1}^{n_g} \text{var}(\log(r_i^2)) + \frac{2}{n_g^2} \sum_{i \in \phi_g, i=1}^{n_g} \sum_{j \in \phi_g, j<i}^{n_g} \rho_{ij}^* \sqrt{\text{var}(\log(r_i^2)) \, \text{var}(\log(r_j^2))}$$

$$= \frac{1}{n_g^2} \sum_{i \in \phi_g, i=1}^{n_g} v_0 + \frac{2}{n_g^2} \sum_{i \in \phi_g, i=1}^{n_g} \sum_{j \in \phi_g, j<i}^{n_g} \rho_{ij}^* v_0$$

Here, ρ_{ij}^* corresponds to correlation between the log-residual-squares $\log(r_i^2)$, for locations *i*, and *j*. Recall that ρ_{ij} is the correlation between r_i and r_j. As ρ_{ij} increases, ρ_{ij}^* too increases, though at a different rate. Further discussions are provided in Section 3.2.2.

Remarks:

1. If the responses are uncorrelated (i.e., $\rho_{ij}^* = 0, i \neq j$), then the second term in (3.6) vanishes, and hence converges with (3.2). Increasing the sample size n_g reduces the variance of the estimate significantly
2. The second term cannot be ignored in presence of correlation. While the first term is the sum of only n_g individual values, the second term is the summation of n_g (n_g - 1) such values. Hence, for considerably high n_g, and high correlation, the second term is significantly larger than the first term. Moreover, increasing n_g does not reduce the second term. Hence, if correlation is moderately high, then increasing the sample size will only reduce the variance of the estimate marginally.

3.2.2 Impact on Variance Modeling

Variance model is obtained by modeling log-variance estimate for each group against the design variables. Section 3.2.1 discussed how correlated response increases the variance of the log-variance estimate. Here, we analyze its impact on the parameter estimation for

variance model obtained through VCVS. First, analytical results are obtained which are complemented by subsequent empirical study.

Consider variance modeling with respect to multiple variables. Variance of the regression parameters is given by (3.5). For a given grouping, it is proportional to $\sigma_{\tilde{s}}^2$. The impact of correlation on the variance is given by the following result.

Proposition 3: Let the design be grouped in κ groups. Let ρ_{ij} and ρ_{ij}^* be defined, as in (12). Then, we have the following result

Let

$$\hat{\sigma}_{\tilde{s}\lg}^2 = \text{var}(\log(\tilde{s}_1^2), \log(\tilde{s}_2^2), ..., \log(\tilde{s}_\kappa^2)) \tag{3.7}$$

Then,

$$E(\hat{\sigma}_{\tilde{s}\lg}^2) = \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\log(\sigma_i^2)^2 + \frac{1}{n_i^2}(\sum_{i=1}^{n_i} v_0 + 2\sum_{i=1}^{n_i}\sum_{j<i}^{n_i}\rho_{ij}^* v_0)] -$$
$$\frac{2}{\kappa(\kappa-1)}\sum_{i=1,}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(l\,\text{og}(\sigma_i^2) + \frac{1}{n_i n_j}\sum_{k\in\phi_i}\sum_{l\in\phi_j} v_0\rho_{kl}^*]$$

Proof: See Appendix

Remarks:

1. As ρ_{ij} (and hence ρ_{ij}^*) increases, both terms in (3.7) increase. Thus, the overall effect is sum total of the effect of correlation on these terms.
2. The first term refers to variance / covariance within group, while the second term refers to between groups. In general, points within group are closer to each other, than those between groups.

3. As discussed before, ρ_{ij}^{*} increases along with ρ_{ij}, but with different rate. This relation is obtained empirically, as seen in Figure 3.5. As ρ_{ij} reduces from 1, ρ_{ij}^{*} reduces much faster than ρ_{ij}. The two values converge at two ends, i.e.,0 and 1.

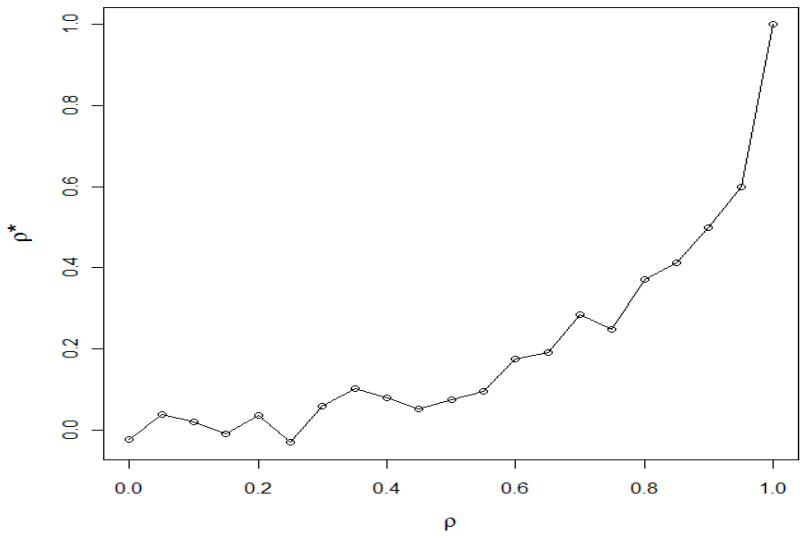


Figure 3.5: Relation between ρ_{ij} and ρ_{ij}^{*}

The effect of correlation on the estimation of variance model is discussed through (3.5) and (3.7) analytically. The analytical results in Sections 3.1 and Section 3.2 have been obtained assuming a particular grouping of design points, and the corresponding variables. Thus, these results correspond to individual tests in Step 2 and Step 3 which test the significance of effects. One can relate the individual tests to the overall performance of VCVS. However, it is difficult to obtain the exact relation because of the way the tests in Step 2 are combined together for variable selection. We leave this for future research. Here, a simulation study is performed in which VCVS is applied to

obtain the variance model. It is assumed that Step 2 of VCVS has identified the variable(s) correctly.

Consider the four models discussed in Section 3.1.2. The following study is performed to investigate the impact of correlation. In the first study (Section 3.2.2.1), we assume the correlation between any two distinct points is the same, i.e., $\rho_{ij} = \rho, i \neq j$. Responses are generated using normal distribution with assumed true variance model and correlation structure. The distribution of the model parameters are obtained by repeating the above analysis 200 times. Correlation coefficient ρ is varied from 0 to 1, and its impact on the result is observed. In the second study (Section 3.2.2.2); we assume a variogram that defines the correlation structure.

3.2.2.1 Constant Correlation

Here, the correlation between two distinct points is assumed to be equal, i.e., $\rho_{ij} = \rho, i \neq j$. The correlation coefficient ρ is increased from 0 to 1. VCVS is applied to estimate the parameters of the variance model. This procedure is repeated 200 times, and the distribution of the model parameters is obtained. The results of this study are illustrated graphically in Figure 3.6 through 3.9. Each figure consist of two graphs illustrating the expected value (left), and the standard error of the estimate (right). The black line with circles denotes the main effect x_1, while the main effect x_2 is represented by a black line. Two-factor-interaction x_1x_2 is represented by blue line with squares denotes while the blue line with cross (x) represents the intercept.

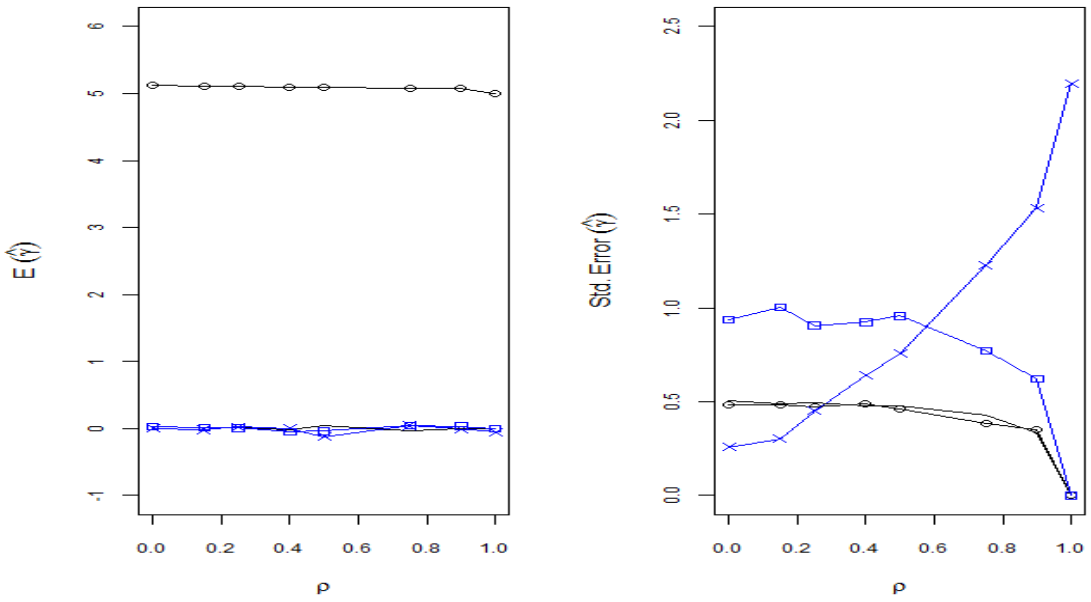


Fig 3.6: Impact of correlation on parameter estimation. Constant correlation.

$\log(\sigma^2) = 5x_1$

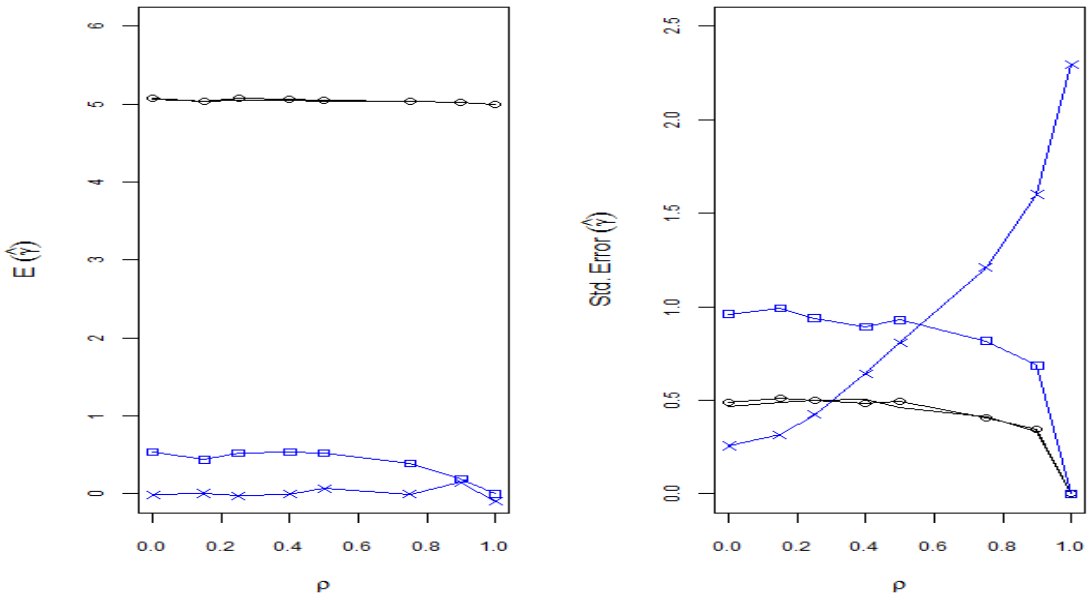


Fig 3.7: Impact of correlation on parameter estimation. Constant correlation.

$\log(\sigma^2) = 5x_1 + 5x_2$

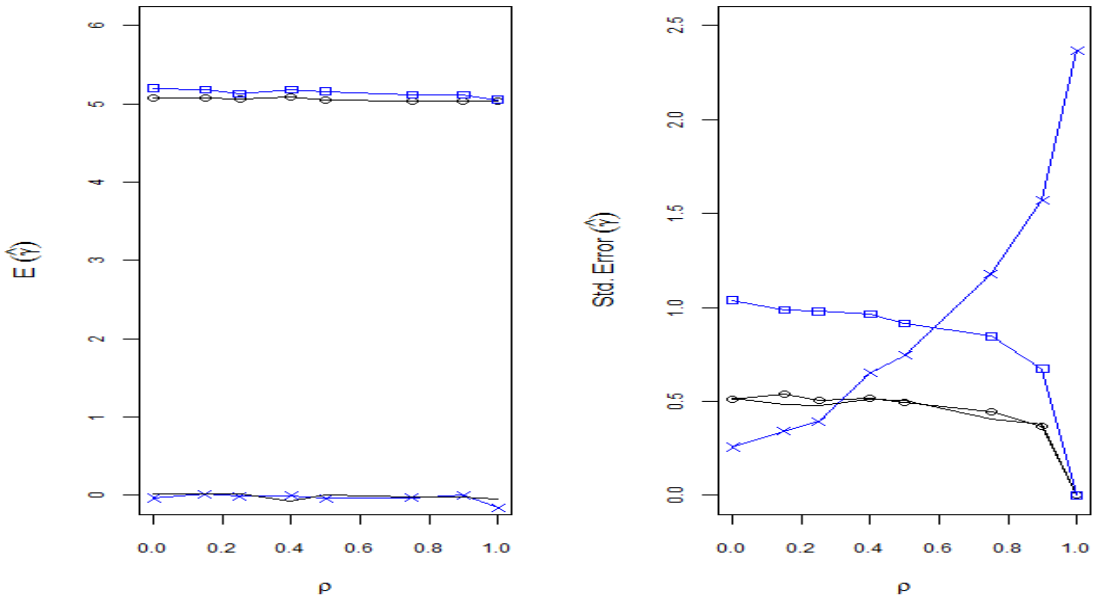


Fig 3.8: Impact of correlation on parameter estimation. Constant correlation. $\log(\sigma^2) = 5x_1+5x_1x_2$

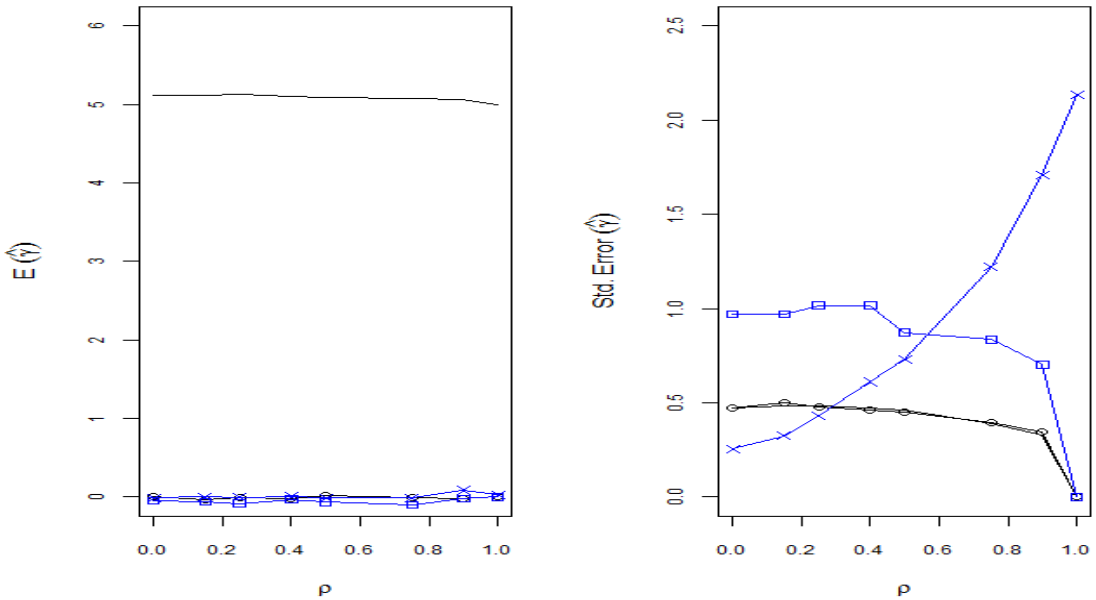


Fig 3.9: Impact of correlation on parameter estimation. Constant correlation. $\log(\sigma^2) = 5x_2$

As discussed before, correlation between the responses does not impact the bias of the parameter estimate. Standard error of the estimate for intercept increases with the correlation. This is because the standard error of log-variance estimate also increases along correlation. However, standard error of the main effect and two-factor- interaction decreases when the correlation is increased. This can be explained through (3.7) as follows. If ρ_{ij} is the same for all distinct points *i, j*, then $\rho_{ij}^{*} = \rho_{kl}^{*}$. Thus, both terms increase significantly with ρ_{ij}. Moreover, as seen from Figure 3.5, ρ_{ij}^{*} increases faster as ρ_{ij} approaches 1 which is reflected in the sharp fall in the standard error of the estimates.

3.2.2.2 Variogram

Here, we analyze the impact of correlation in the form of a variogram. Gaussian correlation structure is assumed i.e., correlation between two points at a distance *h* from each other is given by $R(h) = \exp(-\theta h^2)$. The impact of correlation is measured by varying the correlation parameter θ from 0.01 to 10000. Note that, as θ increases, the correlation between responses *i, j* (ρ_{ij}) decreases. Again, similar to the previous study, VCVS is applied on the data generated from the assumed variance model (four cases) and correlation structure, and repeated 200 times. The results from this study are illustrated graphically in Figure 3.10 through Figure 3.13. For convenience, the correlation parameter θ is displayed in log scale. Note that the correlation between responses increases as we move from left to right, in the graph.

The bias of the estimates is not affected significantly by correlation. However, the correlation affects the standard error of the estimate significantly. As in Section 3.2.2.1, standard error of the intercept increases monotonically with correlation. However, the

impact of correlation on standard error of the main effects and two-factor-interaction effects depends on the true variance model. For example, when the true variance model consists of only one effect (Figure 3.10 and Figure 3.13)), standard error of the inert effects decreases as the correlation is increased. The standard error of active effect increases with correlation until certain point, and then decreases. However, if the variance model consists of more than one active effect, standard error of their inert interaction effect also follows the same pattern as its parent effects, irrespective of its true effect on the variance (Figure 3.11 and Figure 3.12). Results obtained in (3.7) can be used to describe some of these characteristics. In general, the locations within a group are closer than those in different groups. Thus, the between group correlation (ρ_{kl}^{*}) is smaller than within-group correlation (ρ_{ij}^{*}). If correlation between responses ρ_{ij} is low, then the between-group correlation approaches zero. Marginal increase in the first term of (3.7) leads to increase in the standard error. When ρ_{ij} is increases further, it increases the first term significantly, but between-group correlation still remains low, leading to high standard error. However, when ρ_{ij} is high enough, then the between-group correlation increases at a faster rate, as in Figure 3.5. Thus, standard error decreases slightly, as seen in Figures 3.10 to Figure 3.13.

From the above results, it is clear that correlation affects the performance of VCVS. For certain "critical range" of correlation parameter θ, the standard error of the estimates is high. In this study, the design variables are in the range [-1, 1] and the standard errors were high when $\log 10(\theta) \in (-1,1)$. Higher standard errors of active effects and inert effects can result in missing key effects, and identification of spurious

effects respectively. Hence, the proposed VCVS method is not very effective for certain range of correlation between responses.

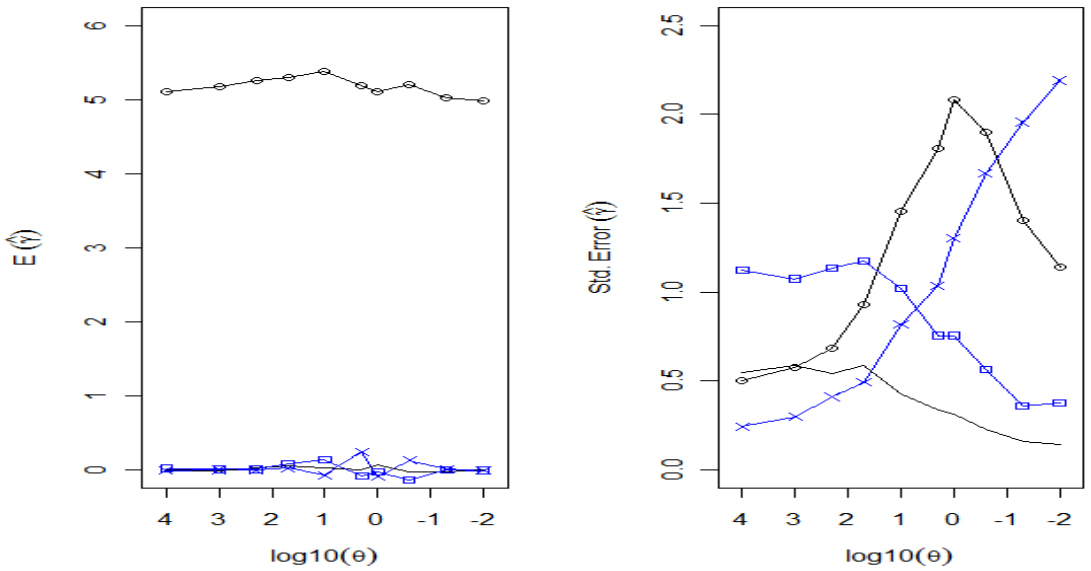


Fig 3.10: Impact of correlation on parameter estimation. Variogram Function. $\log(\sigma^2) = 5x_1$

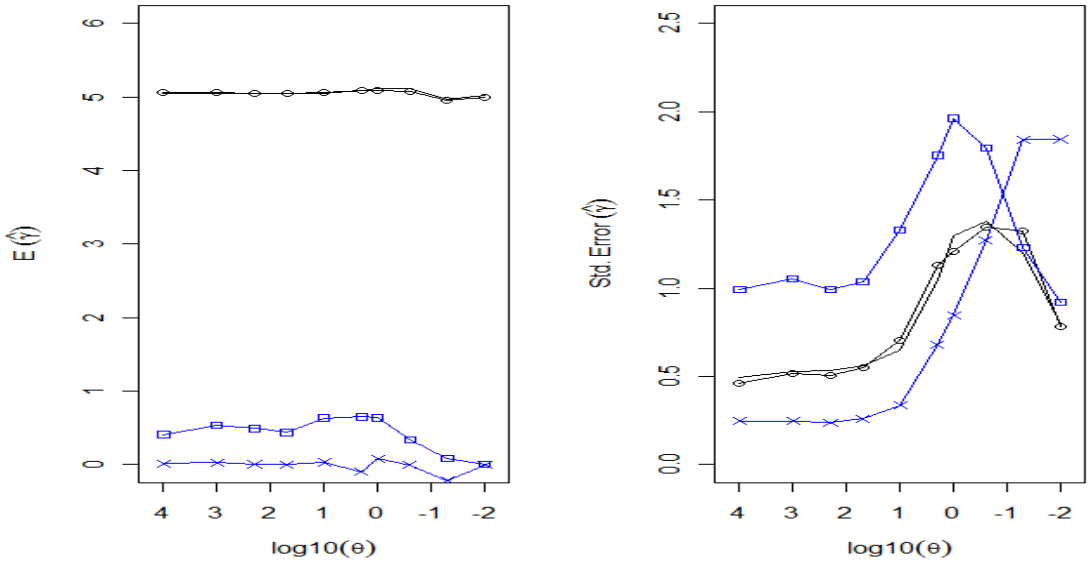


Fig 3.11: Impact of correlation on parameter estimation. Variogram Function. $\log(\sigma^2) = 5x_1+5x_2$

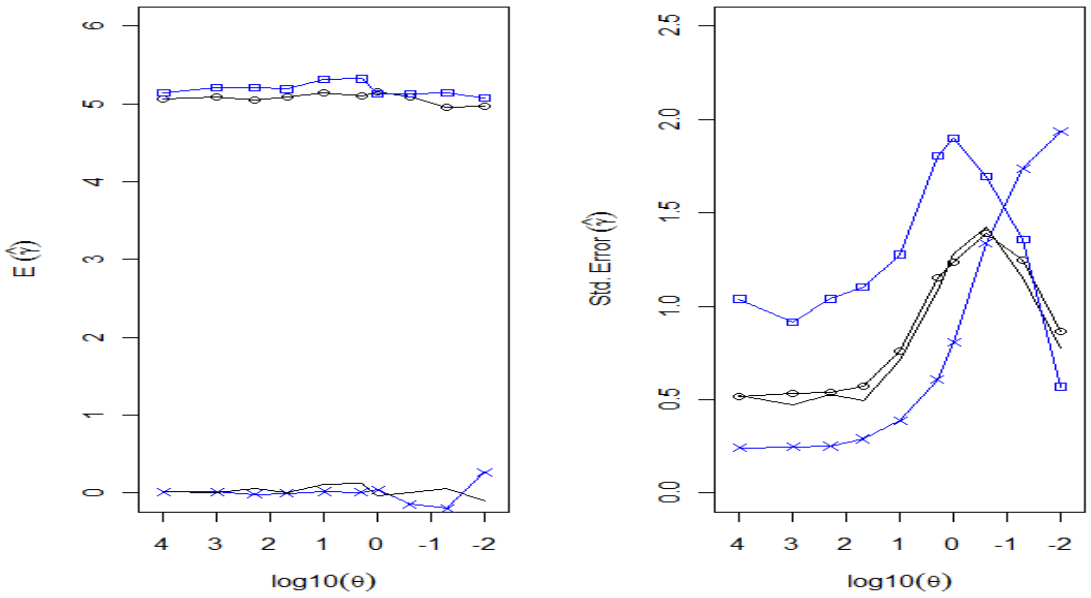


Fig 3.12: Impact of correlation on parameter estimation. Variogram Function. $\log(\sigma^2) = 5x_1+5x_1x_2$

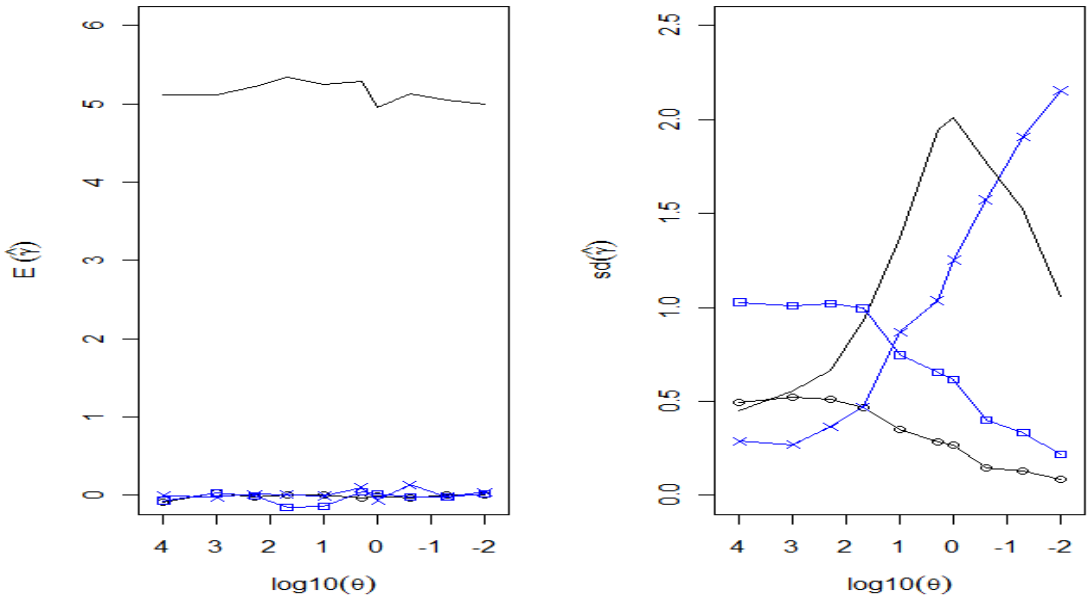


Fig 3.13: Impact of correlation on parameter estimation. Variogram Function. $\log(\sigma^2) = 5x_2$

Remarks:

1. In the above analysis, we assume that the mean model is estimated accurately. Hence, the impact of incorrect mean model on the variance modeling is not reflected in the above results. Since there are numerous studies on this field, we will not go into depth on this topic. Brenneman and Nair (2001), Mcgrath and Lin (2001) etc have researched on this topic.
2. In the next section, VCVS is applied on various hypothetical systems, with linear or non-linear models. The results are then compared with results from variance derivation

3.3. Comparison Study

The performance of VCVS has been analyzed for different situations. Here, we use numerical evaluation methods to compare the performance of VCVS with other methods. VCVS is applied on hypothetical systems with various mean models after the true behavior of the system is approximated by a second order polynomial model. For comparison, TVM is also applied on these systems when applicable. The results from TVM on these systems are compared to those from VCVS. The objective of this study is to identify conditions that favor either of the methods. If the true mean model is non-linear and complex, obtaining true variance model is complicated. In such case, Monte Carlo method can be used to approximate the true variance, as follows.

In RPD, variance is defined as the variation of response with respect to noise variables. Thus, it depends on both distribution of noise, and their relation to the response. Here, noise variables are assumed to follow normal distribution. For each control variable, responses are obtained for numerous (around 200) noise variable

settings, which are chosen randomly from their distribution. The variance thus estimated is modeled against the control variables, to obtain the variance model.

Table 3.1: Comparison study: Estimation of mean model

No	True Model	Estimation of Mean Model	
		Model	R^2_{adj}
1	$y = \exp(1.2x_1z_2)/(x_1+1.2x_2)$	y = 2.1 (Constant)	0
2	$y = \exp(x_1+1.2x_2+z_1)/(x_1+1.2x_2)$	$y = 2 + 3.6x_2$	0.032
3	$y = \cos(x_2z_1)+1.2\sin(x_1z_2)$	$y = 0.85 + 0.25\ x_3^2$	0.04
4	$y = \cos(x_2z_1)$	$y = 1.05 - 0.165x_2^2 - 0.157\ z_1^2$	0.73
5	$y = \exp(1.2x_2+z_2)+\exp(1.2x_2+z_1)$	$y = 1.9 + 1.57x_2 + 1.12x_2^2 + 1.35z_1 + 0.61z_1^2 + 1.41x_2z_1$	0.91
6	$y = \exp(1.2x_1+z_2)$	$y = 1 + 1.55x_1 + 0.8x_1^2 + 1.31N_2 + 0.55z_2^2 + 1.41x_1z_2$	0.98
7	$\mu = 0$, $\log(\sigma^2) = 3.5x_1 + 3x_1x_4$	y = constant	0
8	$y = \exp(x_1+1.2x_2+z_1)/(x_1+0.5z_2)$	$y = 6.4 - 11.4\ x_5^2 - 11.4\ z_2^2$	0.10

Table 3.1 gives details about the true mean model, the estimated model and its accuracy (in terms of R^2_{adj}) in estimating the true model. Table 3.2 gives the details about estimation of variance model, through both methods. It also provides the approximate variance model obtained through Monte Carlo simulation. In case of models 1- 4 and 8, estimated mean model has R^2_{adj} is very small and hence highly inaccurate. While TVM failed to identify any variable that affects the variance model, VCVS is successful in identifying the variance model, at least in part. In case of models 1, 2 and 4, VCVS identified the variables correctly though the models identified might be incorrect. For model 3, VCVS identified only one of the variables in the variance model. Mean model

estimation for models 5, 6 can be considered accurate because the observed R^2_{adj} is high. Both TVM and VCVS successfully identified the variables that affect the variance model.

Table 3.2: Comparison study: Variable selection by TVM and VCVS

No	Monte Carlo Result	Variance through different methods		
		TVM	Variance Modeling VCVS	R^2 of Variance model: R^2
1	$\log(\sigma^2) = 5x_1^2 - 2x_2^2$	None	-5.4x_1x_2	0.85
2	$\log(\sigma^2) = 3 + 2x_1 + 2.6x_2 - x_1^2 - 2x_2^2 - 4x_1x_2$	None	1.4x_2 -4.7x_1x_2	0.73
3	$\log(\sigma^2) = 3x_1^2 + 1.5x_2^2$	None	-2 +1.28x_1	0.65
4	$\log(\sigma^2) = -8 + 9x_2^2\, x_2^2$	None	-5.4+1.2x_2	0.31
5	$\log(\sigma^2) = 2 + 2.4x_2$	x_2	0.4 + 2.45 x_2	0.85
6	$\log(\sigma^2) = 1 + 2.3x_1$	x_1	-0.6 + 3.1x_1	0.82
7	x_1, x_1x_4 : True model	None	3.7x_1+2.8 x_1x_4	0.89
8	$\log(\sigma^2) = 8.5 - 2.1x_1 - 5.31x_1^2 + 2.4x_2$	None	-3.45x_1	0.78

The above analysis uses simpler models like polynomial regression model for estimating the mean model. Appropriate variable selection procedures have been used to obtain parsimonious models. It illustrates how the impact of inaccuracy in estimating mean model impacts the methods for obtaining RPD solution. In practice, advanced modeling methods (e.g. Kriging) can be used for modeling the mean behavior. UP can be applied in such situations. Accurate modeling implies a higher likelihood of variance derived being close to the true variance. However, model 4 in the previous study, and the study by Kunert et al. (2007) counter this argument. Thus it is worth applying the proposed VCVS method even when mean model is estimated accurately.

VCVS method has certain drawbacks. Since the design is split into two based on each variable, it can identify only linear trends of variance, but not the non-linear trend. In example 5, models 3 and 4 in Table 3.1, VCVS identified only the linear effects of the variable. The true model consists of quadratic effects of that variable. In some other cases (models 1 and 2 in Table 3.1), VCVS was not able to identify the true variable. However, in both these cases, the inaccuracy of the estimated model may be the reason behind the poor result of VCVS. It is possible to model the non-linear trend in variance by considering multiple variance-change-points. However, this discussion is beyond the scope of this article. Correlated response was found to increase the standard error significantly, possibly leading to missing key effects, and/or identifying spurious effects. Thus, VCVS can be used for RPD when the estimation of mean model is not very good, and when correlation between responses is low.

3.3.1 Comparison with other Methods

In the previous subsection, the results from VCVS method were compared with that of TVM. It was observed that in general, VCVS method is superior, when the model estimation is not accurate. Here, we discuss the other methods, proposed by Bates et al. (2006), Giovagnoli and Romano (2008), Dellino et al. (2010 and 2011), and Chen et al. (2006) can also be compared with dual modeling. Bates et al. (2006) and Dellino et al. (2011) use the predictions from the metamodels to generate responses and obtain a cross array. Variance is modeled against the control variables. The variance model obtained is heavily impacted by the incorrect mean model because the responses are generated from the estimated model. Similarly, Dellino et al. (2010) obtains the variance expression directly from the mean model (i.e., variance derivation), and hence rely on model

estimation. Chen et al. (2006) proposed UP which is applicable to a variety of metamodels. It expresses the metamodels like spline, kriging etc. as multivariate tensor-product basis functions through which the variance expression can be derived. It can be seen that if model obtained is linear or polynomial, then the resulting variance expression is the same as that obtained from TVM. In comparison to TVM, UP is flexible, as it permits non-linear or complex models. It is rigorous procedure compared to TVM. However, the inaccuracy in model estimation still affects UP significantly. In comparison, incorrect model does not impact the dual modeling approach (VCVS) heavily, as seen from the examples discussed before.

3.4 Application to a Real Life Example: Nanoparticle Synthesis Example

Nanostructures are those with at least one dimension is measured in nanometers (one nanometer (nm) = 10^{-9} meter), hence very small. In comparison, spacing between atoms of a molecule is in the range of 0.12 – 0.15 nm. Hence, properties of many conventional materials change when formed from nanoparticles. Also since nanoparticles have greater surface area per weight, they are more reactive. Nanowires, nanobelt, nanotubes are some examples for nanostructures. Nanostructures have been used in a variety of fields, including but not limited to the following: medicine, structures, sensors, catalysts, and electronics. Nanotubes have novel properties that make them useful in many applications in material science, sensor, optics, architecture, thermal, electrical and electronic appliances.

Here, we consider a system in which Platinum nanoparticles are deposited over carbon nanotubes in supercritical carbon dioxide. The focus is on energy applications, in which transition metals are used as supported catalysts. The experiment considered

consists of two sequential stages. The first step is adsorption stage, and the second step is thermal reduction stage. In the first step, Pt is dissolved in supercritical CO_2 with the help of an organometalic compound (referred to as Preload, or PL). The system is maintained at certain temperature (T) and pressure (P) for some time (Adsorption Time or AdT). Here, Pt compound gets adsorbed on the nanotube. In the second stage, the system is thermally treated for a certain time (Growth time or GrT) during which metallic Pt is deposited on nanotube. The response of interest is the mean size of the Pt nanoparticles. Apart from the factors stated above, the output of the experiment also depends on the number of active sites on the nanotube, referred to CNTLoad. While all the above variables can be controlled, certain factors cannot be controlled. Internal noise (Taguchi (1986)) with respect to temperature (TNoise) and pressure (PNoise), impurity in the raw material (Imp) and Functionalized CNT. Hernandez and Grover (2011) contains a detailed description of this process and the underlying phenomenon through which the process is approximated by a computer simulator.

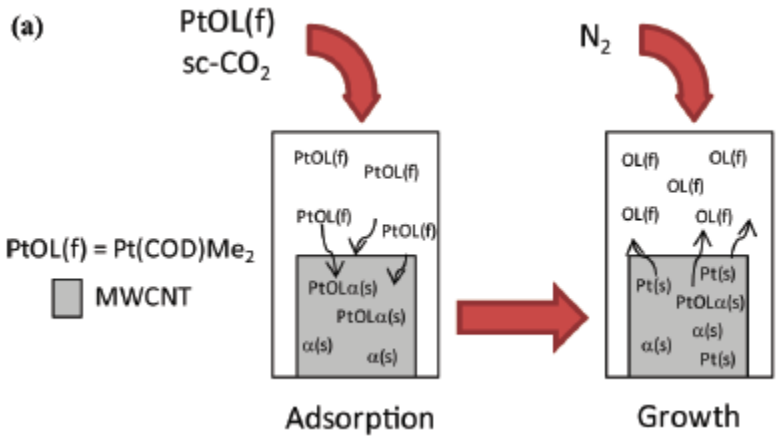


Figure 3.14: Schematic of nanoparticle growth process

The size of the nanoparticles should be controlled for it to be useful in applications like catalysis, medicine, and photonics. Hence, robust parameter optimization is important. Here, since the physical experiment is both time consuming and expensive, they are approximated by computer simulation which takes around three minute. A SFD of size 120 in ten variables (six control and four noise variables) is generated. The simulator is used to obtain response for each setting. Responses are modeled against the predictor variables using a second order polynomial regression model. The parameters of the regression model are provided in Table 3.3. This model has of R^2_{adj} = 0.99. Since the estimated mean model is accurate, TVM is recommended. However, based on the examples in Section 3.3, both TVM and VCVS are expected to give correct results. Both the methods are illustrated for comparison.

VCVS Method: Residuals are obtained using the above estimated model. The estimated mean model is given in Table 3.3. It shows various effects identified and corresponding coefficients. Three control variables T, PL and CNTL have two-factor-interaction effects with one or more noise variables. Hence as per TVM, variance given by the expression below depends on these three variables.

$$\sigma_y^2 = \sigma_{TNoise}^2 (56.6 + 9.5\mathrm{CNTL} - 11.2\mathrm{T})^2 + \sigma_{FCNT}^2 (103 + 14\,\mathrm{PL} + 19.6\mathrm{T} + 14.3\mathrm{CNTL})^2$$

All three variables should be set at their lowest possible value to minimize variance. VCVS is also applied on observed data as follows.

Step 1: VC method is applied for each variable and the results are provided in Table 3.3. T and P do not influence variance because it was observed that the test statistics are low

at most locations, and the peak observed is smaller than the critical value. Hence they will be ignored for the analysis.

Step 2: The following results were obtained in this step.

$M_{PL} = M_{GrT} = M_{AdT}$: None of the effects are significant

M_{CNTL}: $\log(\sigma^2) = 7.06 + 0.75 * CNTL + 0.92 * PL * CNTL$

Hence, the three variables PL and CNTL are considered for step 3

Table 3.3: Estimated mean model for nanoparticle growth process

Effect	Estimate	Effect	Estimate
Intercept	858	T: FCNT	19.6
FCNT	103	FCNT: PreLoad	14
T	80	T:TNoise	-11.2
CNTLoad	71	FCNT: CNTLoad	14.3
TNoise	56.6	T:CNTLoad	11.9
PreLoad	46.7	CNTL: TNoise	9.5
P	-10.3	FCNT: TNoise	9.4
AdT	9.2	CNTL: PreLoad	10
GrT	7.8	$FCNT^2$	-15
GrT^2	-12.7	$Preload^2$	-9
AdT^2	-9.5	$CNTL^2$	-8.5
T^2	-5		

Step 3:

Stepwise regression is performed considering all the effects corresponding to the three variables considered. The variance model is obtained as follows:

$\log(\sigma^2) = 6.3 + 0.8 * CNTL$,

with R^2_{adj} = 0.49

Hence VCVS has identified CNTL in the variance model.

Table 3.4: Change-point for each variable, nanoparticle growth example

Variable	T	P	PL	GrT	AdT	CNTL
G	2.07	1.2	5.67	3.85	7.84	8.93
τ	0.66	0.44	-0.36	-0.56	0.66	0.025

Monte Carlo simulation is employed to approximate the true variance model. Here, control array is a SFD with 80 runs and six control variables, while noise array consists of 200 settings chosen randomly from their distributions. Variance is estimated for each control variables setting, and modeled against the control variables, to obtain the following variance model.

$$\log(\sigma^2) = 7.2 + 0.09\mathrm{T} + 0.22*\mathrm{PL} + 0.94*\mathrm{CNTL} - 0.2*\mathrm{CNTL}^2,$$

Accuracy of the above model is given by R^2_{adj} = 0.79. To summarize the above analyses, estimated mean model was accurate. TVM identified the three variables in the variance model. VCVS has identified the most important variable (CNTL) affecting the variance, while it fails to identify the smaller effects (T, and PL). Both methods rightly determined that CNTL should be minimized to obtain robust setting. Although the predicted variances from the two methods are different, the robust setting with respect to CNTL is the same.

3.5 Concluding Remarks

This research proposes variance-change-point based variable search (VCVS), a new variance modeling methodology for (unreplicated) computer experiment data to support

subsequent robust parameter design studies. VCVS method has been illustrated successfully through application on a nano-particle fabrication simulator.

Variance- change-point concept is used to split the data into two groups, with respect to each control variable. Then, a three-step variable selection procedure identifies variance models. The properties of the proposed method are investigated through analytical and simulation studies of bias and variance of the parameter estimates and also comparisons to variance derivation methods such as Transmitted Variance Model (TVM). When the mean model is estimated accurately, in most cases, both TVM and VCVS lead to correct models. Incorrect mean model impacts both TVM and VCVS, though the impact is lesser on the latter. For the cases that TVM fails to identify the variables influencing the variance, VCVS identifies at least a part of the variance model correctly.

Impact of correlated response on these properties is also analyzed. It was found that correlation affects the variance of estimates, but not bias. For certain level (range) of correlation, there is a possibility of missing key effects, or identifying spurious effects in the variance model. While the impact of correlation on VCVS has been investigated, its impact on methods such as TVM is not known. We leave this for future research.

The method for RPD studies in computer experiments should be based on observed data. This decision depends mainly on accuracy of the estimated mean model, and the correlation between the responses.

VCVS has some limitations. It cannot identify non-linearity in variance model because the data is split into two groups with respect to each variable. Multiple variance-change-points might be employed to identify the non-linearity in the variance. Another

approach is to use other grouping techniques e.g. clustering, nearest neighborhood method, for grouping to create pseudo replicates for variance modeling. Non-linearity in the variance can be identified for clusters being spread over the design space. We leave this study for future research.

Contribution of this chapter

- Investigate the properties of variance modeling procedure, in terms of bias and variance of the estimates
- Analyze the impact of correlated response on the above properties. Illustration of these results through a simulation study.
- Comparison of the proposed variance modeling procedure with TVM, a variance derivation method through some numerical example indicates that the impact of incorrect mean model on the proposed method is lesser than that on TVM.

3.6 References

Bates, R.A., Kenett, R.S., Steinberg, D.M., and Wynn, H.P. (2006). "Achieving Robust Design from Computer Simulations", *Quality Technology and Quantitative Management* 3, 161-177.

Bergman, B., and Hynén, A. (1997), "Dispersion Effects from Unreplicated Designs in the 2^{k-p} Series", *Technometrics*, 39, 191-198.

Box, G.E.P., and Meyer, R.D. (1986), "An Analysis of Unreplicated Fractional Factorials", *Technometrics* 28, 11-18.

Brenneman, W.A., and Nair, V.N. (2001), "Methods for Identifying Dispersion Effects in Unreplicated Factorial Experiments", Technometrics, 43, 388-405.

Chen, J., Gupta, A.K., and Pan, J. (2006a), "Information Criterion and Change Point Problem for Regular Models", *Sankhya: The Indian Journal of Statistics* 68, 252-282.

Chen, W., Jin, R., and Sudjianto, A. (2006), "Analytical Global Sensitivity Analysis and Uncertainty Propagation for Robust Design", *Journal of Quality Technology* 38, 333-348

Cressie, N. (1993). *Statistics for Spatial Data,* Wiley, New York.

Dasgupta, T., Ma, C., Joseph, V.R., Wang, Z.L., and Wu, C.F.J. (2008), "Statistical Modeling for Robust Synthesis of Nanostructures", Journal of the American Statistical Association 103, 594-603.

Dellino, G., Kleijnen, J. P. C. and Meloni, C. (2011), "Robust Optimization in Simulation: Taguchi and Krige Combined ". *INFORMS Journal on computing* doi: 10.1287/ ijoc.1110.0465.

Dellino, G., Kleijnen, J. P. C., and Meloni, C. (2010), "Robust Optimization in Simulation: Taguchi and Response Surface Methodology", *International Journal of Production Economics* 125, 52-59.

Fang, K.T., Li, R., and Sudjianto, A. (2006). *Design and Modeling for Computer Experiments,* CRC Press, New York.

Giovagnoli, A., and Romano, D. (2008), "Robust Design via Simulation Experiments: A Modified Dual Response Approach", *Quality and Reliability Engineering International* 24, 410-416.

Hamada, M., and Wu, C.F.J.(1992), "Analysis of Designed Experiments with Complex Aliasing", *Journal of Quality Technology*, 24, 130-137.

Hawkins, D.M. and Zamba,K.D. (2005), "A Change-Point Model for a Shift in Variance", *Journal of Quality Technology* 37, 21-31.

Hernandez, A.F., and Grover, M.F. (2011), "Comparison of Sampling Strategies for Gaussian Process Models, with Application to Nanoparticle Dynamics", *Industrial and Engineering Chemistry Research* 50, 1379-1388.

Joseph, V.R. (2006), "Limit Kriging". *Technometrics* 48, 458-466.

Joseph, V.R., Hung, Y., and Sudjianto, A. (2008), "Blind Kriging: A New Method for Developing Metamodels", *ASME Journal of Mechanical Design* 130, 1-8.

Kang, L., and Joseph, V.R. (2009), "Bayesian Optimal Single Arrays for Robust Parameter Design", *Technometrics* 51, 250-261.

Kunert, J., Auer, C., and Erdbrugge, M. (2007), "An Experiment to Compare Taguchi's Product Array and the Combined Array", *Journal of Quality Technology,* 39, 17-34.

Mcgrath, R.N., and Lin, D.K.J. (2001), "Confounding of Location and Dispersion Effects in Unreplicated Fractional Factorials", *Journal of Quality Technology* 33, 129-130.

Montgomery, D.C. (1990), "Using Fractional Factorial Designs for Robust Process Development", *Quality Engineering* 3, 193-205

Nair, V.N., and Pregibon, D. (1988), "Analyzing Dispersion Effects from Replicated Factorial Experiments", *Technmetrics*, 30, 247-257.

Pan, G. (1999), "The Impact of Unidentified Location Effects on Dispersion-Effects Identification from Unreplicated factorial Designs", *Technometrics* 41, 313-326.

Robinson, T.J., Borror, C.M., and Myers, R.H. (2004), "Robust Parameter Design: A Review", *Quality and Reliability Engineering International* 20, 81-101.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989), "Design and Analysis of Computer Experiments", *Statistical Science*, 4, 409-423.

Santner, T.J., Williams, B.J., and Notz, W.I. (2003). *The Design and Analysis of Computer Exepriments*. Springer-Verlag, New York.

Shoemaker, A.C., Tsui, K.L., and Wu, C.F.J. (1991), "Economical Experimentation Methods for Robust Design", *Technometrics* 33, 415-427.

Shoemaker, A.C., and Tsui, K.L. (1993), "Response Model Analysis for Robust Design Experiments", *Communication in Statistics – Simulation and Computation* 22, 1037-1064.

Taguchi, G. (1986). *Introduction to Quality Engineering*. Kraus international publications, White Plains, NY.

Vining, G.G., and Myers, R.H. (1990), "Combining Taguchi and response surface philosophies: A dual response approach" *Journal of Quality Technology* 22, 38-45

Welch, W.J., Yu, T.K., Kang, S.M., and Sacks, J. (1990), "Computer Experiments for Quality Control by Parameter Design", *Journal of Quality Technology* 22, 15-22

Woo, H. (2010), Ph.D. Thesis Proposal Report, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, (advisor: Drs. J-C Lu, B. Vidakovic and M. Grover).

Wu, C.F.J., and Hamada, M. (2009), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York, Wiley, 2nd edition.

Xu, S., Adiga, N., Ba, S., Dasgupta, T., Wu, C.F.J., and Wang, Z.L. (2009), "Optimizing and Improving the Growth Quality of ZnO Nanowire Arrays Guided by Statistical Design of Experiments", *American Chemical Society Nano* 3, 1803-1812.

APPENDIX A

PROOF OF RESULTS IN CHAPTER 1

Proof of (1.2). If VS identifies only one factor as active, it does not involve any capping run (a minimum of two active factors is necessary to conduct a capping run). However, each of the k factors undergoes swapping, which results in a total of $2k$ runs. Adding the 6 stage-I runs to this, we have $N = 2(k + 3)$. For $m > 2$, the total number of capping runs will be $2(m-1)$, and the total number of swapping terms will range from $2m$ to $2k$ depending on the order of swapping. Again, the result follows by adding the six stage I runs to the number of swapping and capping runs.

Proof of Theorem 1. From assumption (i), it follows that the total number of active factors identified by VS will be exactly p. By (1.2), N can take integer values ranging from $4(p+1)$ to $2(k+p)+4$. Clearly, N varies *only* due to the number of extra swapping runs, which equals $2j$ if j inactive factors are examined, $j = 0, 1, \ldots, (k-p)$. Therefore, N can take only even integer values $4(p + 1), 4(p + 1) + 2, \ldots, 4(p + 1) + 2(k - p)$.

The total number of mutually exclusive, exhaustive and equally likely ways in which k factors can be examined is $k!$. The event $\mathcal{A}_j = \{N = 4(p + 1) + 2j\}$ would occur if the search contains j extra swaps, i.e., j inactive factors have to be explored before the pth active factor. This will happen if the $(p + j)$th factor examined is active, and there are j inactive factors among the first $(p + j - 1)$ factors examined. The total possible number of such arrangements is given by

$$n(\mathcal{A}_j) = \binom{p}{1}\binom{k-p}{j}\binom{p+j-1}{j}(p-1)!\;\; j!\;\; (k-p-j)!.$$

Since each of the above arrangements are mutually exclusive, exhaustive and equally

likely, by classical definition of probability, it follows that

$$Pr(\mathcal{A}_j) = \frac{n(\mathcal{A}_j)}{k!} = \frac{\binom{p}{1}\binom{k-p}{j}\binom{p+j-1}{j}(p-1)!\ \ j!\ \ (k-p-j)!}{k!},$$

which, after some algebraic manipulations, can be written in a more convenient form $Pr(\mathcal{A}_j) = p\binom{k-p}{j}/(p+j)\binom{k}{p+j}$

The expectation of N is given by

$E(N) = \sum_{j=0}^{k-p}\Big(4(p+1)+2j\Big)\ \ Pr(\mathcal{A}_j)$

$= 4(p+1) + 2\sum_{j=1}^{k-p} j\ \ Pr(\mathcal{A}_j)\ \ (\text{since}\ \ \sum_{j=0}^{k-p} Pr(\mathcal{A}_j) = 1)$

$= 4(p+1) + 2\sum_{j=1}^{k-p} j\ \ \frac{p}{p+j}\frac{\binom{k-p}{j}}{\binom{k}{p+j}} = 4(p+1) + 2\frac{p}{p+1}(k-p)$

after some algebraic manipulations.

Proof of Lemma 1.

Let x_{il} denote the lth element of column $\mathbf{x}_i$, $i = 1, \ldots, p$, $l = 1, 2, \ldots, 2p+2$, which takes value either -1 or +1. The lth element of the column $\mathbf{z}_{ij}$ that represents the interaction between factors i and j is given by $z_{ijl} = x_{il}x_{jl}$. We shall prove that $\mathbf{z}'_{ij}\mathbf{x}_i = 0$, i.e., $\sum_{l=1}^{2p+2} z_{ijl}x_{il} = 0$. Now, $\sum_{l=1}^{2p+2} z_{ijl}x_{il} = \sum_{l=1}^{2p+2} x_{il}^2 x_{jl} = \sum_{l=1}^{2p+2} x_{jl}$

Note that, in each block of 2 runs of the VS design, the settings can be obtained by swapping "+" and "-" signs. Due to this, sum of x_j for any j in any block is equal to 0. Hence $\sum_{l=1}^{2p+2} x_{jl} = 0$

Proof of Theorem 2.

(i): Let $\mathbf{Z}_p$ denote the $(2p+2) \times \binom{p}{2}$ matrix obtained by taking the pairwise product of the p columns of the $\mathbf{X}_p$ matrix. Also, let $\boldsymbol{\beta}_{INT}$ denote the $\binom{p}{2} \times 1$ vector of coefficients β_{ij}'s defined in model (1.3). Assuming absence of 3fi's and higher order interactions, from model (1.3) we can write $E(\mathbf{y}) = \beta_0\mathbf{1}_{2p+2} + \mathbf{X}_p\boldsymbol{\beta}_{main} + \mathbf{Z}_p\boldsymbol{\beta}_{INT}$, where $\mathbf{1}_N$ denotes the $N \times 1$ vector of 1's. Now, by definition of $\hat{\boldsymbol{\beta}}_{main}$,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{main}) &= E\big[(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{y}\big] \\ &= E\big[(\mathbf{X}'_p\mathbf{X}_p)^{-1}(\mathbf{X}'_p\beta_0\mathbf{1}_{2p+2} + \mathbf{X}'_p\mathbf{X}_p\boldsymbol{\beta}_{main} + \mathbf{X}'_p\mathbf{Z}_p\boldsymbol{\beta}_{INT})\big] = \boldsymbol{\beta}_{main}. \end{aligned}$$

because $\mathbf{X}_p'\mathbf{1}_{2p+2} = \mathbf{0}$ and by Lemma 1, $\mathbf{X}_p'\mathbf{Z}_p = \mathbf{0}$.

(ii): The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{main}$ is $\sigma^2\mathbf{D}_p^{-1}$, where $\mathbf{D}_p = (\mathbf{X}_p'\mathbf{X}_p)$. It is easy to verify that

$$\mathbf{D}_p = \begin{bmatrix} a_p & b_p & b_p & \cdots & b_p \\ b_p & a_p & b_p & \cdots & b_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_p & b_p & b_p & \cdots & a_p \end{bmatrix}, \tag{A.1}$$

where $a_p = 2p + 2$ and $b_p = 2p - 6$ are functions of p. The determinant of $\mathbf{D}_p$ can be obtained as:

$$\det \mathbf{D}_p = \big[a_p + (p-1)b_p\big](a_p - b_p)^{p-1}. \tag{A.2}$$

Let $\mathbf{D}_p^r$ denote the $r \times r$ principal submatrix of $\mathbf{D}_p$ for $r = 1, \ldots, p$. Then,

$$\det \mathbf{D}_p^r = \big[a_p + (r-1)b_p\big](a_p - b_p)^{r-1}. \tag{A.3}$$

Clearly, all the diagonal elements of $\mathbf{D}_p^{-1}$ will be the same and, when multiplied by σ^2, will represent $var(\hat{\beta}_i)$ for $i = 1, \ldots, p$. In particular, the $(1,1)$th element of $\mathbf{D}_p^{-1}$ will be equal to the adjugate of the $(1,1)$th element of $\mathbf{D}_p$ divided by $\det \mathbf{D}_p$. Since the adjugate of the $(1,1)$th element of $\mathbf{D}_p$ is $\det \mathbf{D}_p^{p-1}$, it follows that

$$\begin{aligned} var(\hat{\beta}_i) &= \sigma^2 \frac{\det \mathbf{D}_p^{p-1}}{\det \mathbf{D}_p} = \sigma^2 \frac{\big[a_p + (p-2)b_p\big]}{\big[a_p + (p-1)b_p\big](a_p - b_p)}, \quad \text{by (A.2) and (A.3)} \\ &= \sigma^2 \frac{p^2 - 4p + 7}{8(p^2 - 3p + 4)}. \end{aligned}$$

Proceeding in the same way, it can be shown that any off-diagonal element of $\mathbf{D}_p^{-1}$ can be written as $(\det \mathbf{C}_{p-1})/(\det \mathbf{D}_p)$, where

$$\mathbf{C}_p = \begin{bmatrix} b_p & b_p & b_p & \cdots & b_p \\ b_p & a_p & b_p & \cdots & b_p \\ b_p & b_p & a_p & \cdots & b_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_p & b_p & b_p & \cdots & a_p \end{bmatrix}, \tag{A.4}$$

and a_p and b_p are defined as before. Since $\det \mathbf{C}_{p-1} = -b_p(a_p - b_p)^{p-2} = -8^{p-2}(2p-6)$, we have $cov(\hat{\beta}_i, \hat{\beta}_j) = -\frac{p-3}{8(p^2-3p+4)}\sigma^2$, by substituting the values of a_p and b_p.

Proof of Theorem 3.

Part (i)

Consider any three columns of $\mathbf{D}$ that correspond to factors i, j and k identified as active. Then, the rows corresponding to stage-1 runs are $(+,+,+)$ and $(-,-,-)$. It is straightforward to verify that the swapping runs for factors i, j and k will generate the six remaining combinations of a 2^3 design.

Parts (ii)–(iv)

It is easy to see that the 8 swapping runs for any four factors A, B, C and D generate a 2^{4-1} design with the defining relation $I = -ABCD$. Note that if an additional column is introduced, six rows of that column will have the same symbol ($-$ or $+$) as in one of the four columns. Hence, the augmented matrix will not be an orthogonal array. Next, we show that $\mathbf{D}$ cannot contain an orthogonal array with more than 8 rows. Let $i_1, i_2, \ldots i_p$ denote the p active factors in the order in which they are identified as active. For each pair (i_k, i_{k+1}), the level combinations $(-,+)$ and $(+,-)$ can appear *twice and only twice* corresponding to the two swapping runs. Thus, the largest orthogonal submatrix of the $N \times 2$ matrix with columns i_k and i_{k+1} has 8 rows. To prove the last part, we note that the four runs along with two stage-I runs and two capping runs $(+,+,-,-)$ and $(-,-,+,+)$ form a 2^{4-1} design with the defining relation $I = ABCD$. Therefore, the combination of the two half-fractions with defining relations $I = -ABCD$ and $I = ABCD$ constitute a 2^4 full factorial design.

Hypothesis test for Stage I and derivation of (1.13) and (1.14).

Since the variance of the median obtained from a random sample of size n drawn

from a normal population can be approximated as $\pi/2n$ for large n, it is easy to see that the distribution of $M_b - M_w$ can be approximated as $N(2\sum_{i=1}^{k}\beta_i, (\pi\sigma^2)/3)$. Now, let $\hat{\sigma_b}^2$ and $\hat{\sigma_w}^2$ denote the sample variances of the two sets of observations on y^+ and y^- respectively and $\hat{\sigma}^2 = (\hat{\sigma_b}^2 + \hat{\sigma_w}^2)/2$ denote the pooled variance. Then, $(M_b - M_w)/(\hat{\sigma}\sqrt{\pi/3})$ follows a non-central t distribution with 4 degrees of freedom and non-centrality parameter $\delta = (2\sum_{i=1}^{k}\beta_i)/(\sigma\sqrt{\pi/3})$. Note that $\sum_{i=1}^{k}\beta_i > 0$ implies at least one of the factors is active.

Hypothesis test for the swapping stage and derivation of (1.15).

Define the following statistics with respect to the swapping of factor x_i:

$$\begin{aligned} s^-(x_i) &= M_b - y_i^-, \\ s^+(x_i) &= y_i^+ - M_w. \end{aligned} \tag{A.5}$$

Since M_b is distributed approximately as $N(\beta_0 + \sum_j \beta_j + \sum_{j\neq k}\beta_{jk}, (\pi/6)\sigma^2)$ and $y_i^- \sim N(\beta_0 - \beta_i + \sum_{j\neq i}\beta_j - \sum_{j\neq i}\beta_{ij} + \sum_{j\neq i,k\neq i}\beta_{jk}, \sigma^2)$, the distribution of $s^-(x_i)$ can be approximated by $N(2(\beta_i + \sum_{j\neq i}\beta_{ij}), (1+\pi/6)\sigma^2)$. Similarly, the distribution of $s^+(x_i)$ can be approximated by $N(2(\beta_i - \sum_{j\neq i}\beta_{ij}), (1+\pi/6)\sigma^2)$. Note that factor x_i is active if and only if at least one of the following conditions holds good: (i) $\beta_i \neq 0$ (ii) $\sum_{j\neq i}\beta_{ij} \neq 0$. Thus the following null hypothesis of interest at this stage is $H_0 : \beta_i = \sum_{j\neq i}\beta_{ij} = 0$. The hypothesis is rejected at level α if either or both of the following two events occur:

$$\begin{aligned} A^+ &: \frac{|s^+(x_i)|}{1.23\hat{\sigma}} > t_{4,\alpha/2}, \\ A^- &: \frac{|s^-(x_i)|}{1.23\hat{\sigma}} > t_{4,\alpha/2}, \end{aligned} \tag{A.6}$$

where $1.23 = (1+\pi/6)^{1/2}$. Since A^+ and A^- are independent events, (1.15) follows immediately.

Hypothesis test for the capping stage and derivation of (1.18).

Define the following statistics with respect to the capping of factors in $\mathcal{F}$:

$$\begin{aligned} C^+(\mathcal{F}) &= M_b - y^+_{\mathcal{F}}, \\ C^-(\mathcal{F}) &= y^-_{\mathcal{F}} - M_w. \end{aligned} \tag{A.7}$$

Proceeding as before, it can be easily seen that

$$\begin{aligned} C^+(\mathcal{F}) &\sim N(2\sum_{i\notin\mathcal{F}}\beta_i + 2\sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij}, (1+\pi/6)\sigma^2), \\ C^-(\mathcal{F}) &\sim N(2\sum_{i\notin\mathcal{F}}\beta_i - 2\sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij}, (1+\pi/6)\sigma^2). \end{aligned}$$

Note that the capping runs will be successful (i.e., all factors $x_i, i \notin \mathcal{F}$ will be declared inert) if and only if both the following conditions hold good: (i) $\sum_{i\notin\mathcal{F}}\beta_i = 0$ and (ii)$\sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij} = 0$. Thus the following null hypothesis of interest at this stage is $H_0 : \sum_{i\notin\mathcal{F}}\beta_i = \sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij} = 0$. The hypothesis is rejected (i.e., the capping run declared unsuccessful) at level α if either or both of the following two events occur:

$$\begin{aligned} B^+ &: \frac{|C^+(\mathcal{F})|}{1.23\hat{\sigma}} > t_{4,\alpha/2}, \\ B^- &: \frac{|C^-(\mathcal{F})|}{1.23\hat{\sigma}} > t_{4,\alpha/2}. \end{aligned} \tag{A.8}$$

Since B^+ and B^- are independent events, (1.18) follows immediately.

Hypotheses testing in the presence of three-factor interactions.

Due to the presence of three-factor interactions, the expectations of the test statistics associated with stage I, swapping and capping require the following modifications.

$$\begin{aligned} E(M_b - M_w) &= 2(\sum_{i=1}^{k}\beta_i + \sum_{i<j,l}\sum_{j<l}\sum_{l=1}^{k}\beta_{ijl}), \\ E(s^+(x_i)) &= 2(\beta_i + \sum_{j\neq i}\sum_{l\neq i,j}\beta_{ijl} + \textstyle\sum_{j\neq i}\beta_{ij}), \\ E(s^-(x_i)) &= 2(\beta_i + \sum_{j\neq i}\sum_{l\neq i,j}\beta_{ijl} - \textstyle\sum_{j\neq i}\beta_{ij}), \\ E(C^+(\mathcal{F})) &= 2\sum_{i\notin\mathcal{F}}\beta_i + 2\sum_{\psi(\mathcal{F})}\sum\sum\beta_{ijl} + 2\sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij}, \\ E(C^-(\mathcal{F})) &= 2\sum_{i\notin\mathcal{F}}\beta_i + 2\sum_{\psi(\mathcal{F})}\sum\sum\beta_{ijl} - 2\sum_{i\notin\mathcal{F}}\sum_{j\in\mathcal{F}}\beta_{ij}, \end{aligned}$$

where $\psi(\mathcal{F})$ is the set of all triplets (i, j, l) where one or all of the $i, j, l \notin \mathcal{F}$.

The variance expressions and the test procedures are exactly identical to those derived in the Appendices before.

APPENDIX B

RESULTS FOR CHAPTER 2 AND CHAPTER 3

1. Proof of Proposition 1.

Proof:

$r_i \sim N(0,\sigma_i^2)$. Hence, $E(r_i^2) = E(r_i)^2 + \text{var}(r_i) = 0 + \sigma_i^2 = \sigma_i^2$

Also, note that $\frac{r_i^2}{\sigma_i^2} \sim \chi_1^2$. Hence $r_i^2 = \sigma_i^2 \chi_1^2$

Taking logarithm of the above,

$$\log(r_i^2) = \log(\sigma_i^2) + \log(\chi_1^2)$$

Taking expectation over both sides,

$$E(\log(r_i^2)) = E(\log(\sigma_i^2)) + E(\log(\chi_1^2))$$

Hence, $\log(r_i^2) - \log(\chi_1^2)$ is an unbiased estimator of $\log(\sigma_i^2)$

$$E(\log(\chi_k^2)) = \log(2) + \psi(k/2),$$

where Ψ is digamma function, defined as follows.

$$\psi(z) = \frac{d}{dz}\ln\Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$$

$\psi(1/2) = -\varsigma - 2\log(2)$, where ς is Euler – Mashceroni constant ≈ 0.5772

Hence, $E(\log(\chi_1^2)) = \log(2) - \varsigma - 2\log(2) = -\varsigma - \log(2) \approx$ -1.27036

Therefore, $\log(r_i^2) + \log(\eta)$ is unbiased estimator of $\log(\sigma_i^2)$, where

$\eta = \exp(-E(\log(\chi_1^2))) \approx \exp(1.27)$

Remarks: Proposition 1 justifies the factor η in estimating variance $\tilde{s}^2 = \eta(\prod_{i=1}^{n} r_i^2)^{1/n}$

2. Proof of Proposition 2.

$$E(\hat{\gamma}_1) = \frac{E(\frac{1}{n_2}\sum_{i\in\phi_2, i=1}^{n_2}\log(r_i^2) - \frac{1}{n_1}\sum_{i\in\phi_i i=1}^{n_1}\log(r_i^2))}{\mu_2 - \mu_1}$$

where μ_1, μ_2 are the mean x_1 values for groups 1 and 2 respectively.

$$= \frac{\frac{1}{n_2}(\sum_{i\in\phi_2, i=1}^{n_2}\log(\sigma_i^2) + E(\chi_1^2)) - \frac{1}{n_1}(\sum_{i\in\phi_1, i=1}^{n_1}\log(\sigma_i^2) + E(\chi_1^2))}{\mu_2 - \mu_1}$$

$$= \frac{\frac{1}{n_2}\sum_{i\in\phi_2, i=1}^{n_2}\log(\sigma_i^2) - \frac{1}{n_1}\sum_{i\in\phi_1, i=1}^{n_1}\log(\sigma_i^2)}{\mu_2 - \mu_1}$$

(From 2.2, $\log(\sigma^2) = \gamma_1 x$)

$$= \frac{\frac{1}{n_2}\sum_{i\in\phi_2, i=1}^{n_2}\gamma_1 x_i - \frac{1}{n_1}\sum_{i\in\phi_1, i=1}^{n_1}\gamma_1 x_i}{\mu_2 - \mu_1} = \frac{\frac{1}{n_2}\gamma_1\sum_{i\in\phi_2, i=1}^{n_2} x_i - \frac{1}{n_1}\gamma_1\sum_{i\in\phi_1, i=1}^{n_1} x_i}{\mu_2 - \mu_1}$$

$$= \frac{\frac{1}{n_2}\gamma_1(n_2\mu_2) - \frac{1}{n_1}\gamma_1(n_1\mu_1)}{\mu_2 - \mu_1} = \gamma_1$$

The above result is valid for one variable case, with two groups. Consider the multiple variable case, with the data divided into numerous groups. Let **X** be design matrix, as in (3.5) which is divided into κ groups. Denote the vector of estimates for log-variance for each of the groups as follows:

$$\tilde{\mathbf{s}}_{\text{lg}} = (\log(\tilde{s}_1^2), \log(\tilde{s}_2^2), \ldots, \log(\tilde{s}_\kappa^2))$$

Then the regression coefficients for the log-variance model (2.2) is given by

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\tilde{\mathbf{s}}_{\text{lg}}$$

Thus, the expected value of this statistic is given by

$$E(\hat{\boldsymbol{\gamma}}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{s}}_{\text{lg}})$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\tilde{\mathbf{s}}_{\text{lg}})$$

Now,

$$E(\tilde{\mathbf{s}}_{\text{lg}}) = (E(\log(\tilde{s}_1^2)), E(\log(\tilde{s}_{12}^2)), \ldots, E(\log(\tilde{s}_\kappa^2))$$

$$= (\log(\sigma_1^2), \log(\sigma_2^2), \ldots, \log(\sigma_\kappa^2))$$

$$= \mathbf{X}\boldsymbol{\gamma}$$

Thus,

$$E(\hat{\boldsymbol{\gamma}}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\gamma})$$

$$= \boldsymbol{\gamma}$$

3. Proof of Proposition 3

$$\hat{\sigma}^2_{\tilde{s}\lg} = \text{var}(\log(\tilde{s}_1^2), \log(\tilde{s}_2^2), ..., \log(\tilde{s}_\kappa^2))$$

Second central moment for the above vector is given by

$$\frac{1}{\kappa}\sum_{i=1}^{\kappa}\log(\tilde{s}_i^2)^2 - (\frac{1}{\kappa}\sum_{i=1}^{\kappa}\log(\tilde{s}_i^2))^2$$

(After expansion, and some simplifications)

$$= \frac{\kappa-1}{\kappa^2}\sum_{i=1}^{\kappa}\log(\tilde{s}_i^2)^2 - \frac{2}{\kappa^2}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}\log(\tilde{s}_i^2)\log(\tilde{s}_j^2)$$

Hence

Sample variance $\hat{\sigma}^2_{\tilde{s}\lg}$ is given by

$$\hat{\sigma}^2_{\tilde{s}\lg} = \frac{\kappa}{\kappa-1}[\frac{\kappa-1}{\kappa^2}\sum_{i=1}^{\kappa}\log(\tilde{s}_i^2)^2] - \frac{\kappa}{\kappa-1}[\frac{2}{\kappa^2}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}\log(\tilde{s}_i^2)\log(\tilde{s}_j^2)]$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}\log(\tilde{s}_i^2)^2 - \frac{2}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}\log(\tilde{s}_i^2)\log(\tilde{s}_j^2)]$$

Thus,

$$\text{E}(\hat{\sigma}^2_{\tilde{s}\lg}) = \frac{1}{\kappa}\sum_{i=1}^{\kappa}E(\log(\tilde{s}_i^2)^2) - \frac{2}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}E(\log(\tilde{s}_i^2)\log(\tilde{s}_j^2))$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[E(\log(\tilde{s}_i^2))^2 + \text{var}(\log(\tilde{s}_i^2))]$$

$$- \frac{2}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}[E(\log(\tilde{s}_i^2))E(\log(\tilde{s}_j^2)) + \text{cov}(\log(\tilde{s}_i^2), \log(\tilde{s}_j^2))]$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[E(\log(\tilde{s}_i^2))^2 + \mathrm{var}(\log(\tilde{s}_i^2))]$$
$$-\frac{2}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(\log(\sigma_j^2) + \mathrm{cov}(\log(\tilde{s}_i^2), \log(\tilde{s}_j^2))]$$

(Using the results obtained in Section 3.1, Section 3.2.1)

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\log(\sigma_i^2)^2 + \mathrm{var}(\log(\tilde{s}_i^2))]$$
$$-\frac{2}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(l\,\mathrm{og}(\sigma_i^2) + \mathrm{cov}(\frac{1}{n_i}\sum_{k\in\phi_i}\log(r_k^2), \frac{1}{n_j}\sum_{l\in\phi_j}\log(r_l^2))]$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\log(\sigma_i^2)^2 + \mathrm{var}(\log(\tilde{s}_i^2))]$$
$$-\frac{2}{\kappa(\kappa-1)}\sum_{i=1,}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(l\,\mathrm{og}(\sigma_i^2) + \frac{1}{n_i n_j}\sum_{k\in\phi_i}\sum_{l\in\phi_j}\rho_{kl}^*\sqrt{\mathrm{var}(\log(r_k^2))\,\mathrm{var}(\log(r_l^2))}]$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\log(\sigma_i^2)^2 + \mathrm{var}(\log(\tilde{s}_i^2))]$$
$$-\frac{2}{\kappa(\kappa-1)}\sum_{i=1,}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(l\,\mathrm{og}(\sigma_i^2) + \frac{1}{n_i n_j}\sum_{k\in\phi_i}\sum_{l\in\phi_j}v_0\rho_{kl}^*]$$

$$= \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\log(\sigma_i^2)^2 + \frac{1}{n_i^2}(\sum_{i=1}^{n_i}v_0 + 2\sum_{i=1}^{n_i}\sum_{j<i}^{n_i}v_0\rho_{ij}^*]$$
$$-\frac{2}{\kappa(\kappa-1)}\sum_{i=1,}^{\kappa}\sum_{j<i}^{\kappa}[\log(\sigma_i^2)(l\,\mathrm{og}(\sigma_i^2) + \frac{1}{n_i n_j}\sum_{k\in\phi_i}\sum_{k\in\phi_j}v_0\rho_{kl}^*]$$

VITA

Nagesh Adiga was born in India. He received a Bachelor's degree in Mechanical Engineering from Karnataka Regional Engineering College (Now known as National Institute of Technology Karnataka), Surathkal. He worked for one year as graduate engineer in Larsen and Tourbo Limited, Hazira, Surat before pursuing M.Tech (Quality, Reliability and Operations Research) from Indian Statistical Institute, Kolkata, India. He joined the School of Industrial and Systems Engineering at Georgia Institute of Technology as a doctoral student in 2006.